Structural Optimization for Large-Scale Problems

Lecture 1: Intrinsic complexity of Black-Box Optimization

Yurii Nesterov

Minicourse: November 15, 16, 22, 23, 2024 (SDS, Shenzhen)

Outline

Basic NP-hard problem

NP-hardness of some popular problems

Lower complexity bounds for Global Minimization

Nonsmooth Convex Minimization. Subgradient scheme.

Smooth Convex Minimization. Lower complexity bounds

Methods for Smooth Minimization with Simple Constraints

Standard Complexity Classes

Let data be coded in matrix A, and n be dimension of the problem.

Combinatorial Optimization

- NP-hard problems: 2^n operations. Solvable in $O(p(n)\|A\|)$.
- Fully polynomial approximation schemes: $O\left(\frac{p(n)}{\epsilon^k}\ln^\alpha \|A\|\right)$.
- Polynomial-time problems: $O(p(n)\ln^\alpha \|A\|)$.

Continuous Optimization

- Sublinear complexity: $O\left(\frac{p(n)}{\epsilon^\alpha}\|A\|^\beta\right)$, $\alpha, \beta > 0$.
- Polynomial-time complexity: $O\left(p(n)\ln(\frac{1}{\epsilon}\|A\|)\right)$.

Basic NP-hard problem: Problem of stones

1.2. NP-hard 问题的例子

e.g. 1

Given n stones of integer weights $a_1, \ldots, a_n$, decide if it is possible to divide them on two parts of equal weight.

Mathematical formulation

Find a Boolean solution $x_i = \pm 1$, $i = 1, \ldots, n$, to a single linear equation

$$\sum_{i=1}^{n} a_i x_i = 0.$$

Another variant: $\sum_{i=2}^{n} a_i x_i = a_1.$ NP-hard

NB: Solvable in $O\left(\ln n \cdot \sum_{i=1}^{n} |a_i|\right)$ by FFT transform.

Immediate consequence: quartic polynomial

Theorem: Minimization of quartic polynomial of n variables is NP-hard.

Proof: Consider the following function:

0是 minimum

$$f(x) \quad = \quad \sum_{i=1}^{n} x_i^4 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right)^2 + \left(\sum_{i=1}^{n} a_i x_i\right)^4 + (1 - x_1)^4.$$

The first part is $\langle A[x]^2, [x]^2 \rangle$, where $A = I - \frac{1}{n} e_n e_n^T \succeq 0$ with $Ae_n = 0$, and $[x]_i^2 = x_i^2$, $i = 1, \ldots, n$.

$A'y=0$ 为 non-singular

全是1的 $n \times n$ 矩阵

为唯一解

Thus, $f(x) = 0$ iff all $x_i^2 = \tau$, $\sum_{i=1}^{n} a_i x_i = 0$, and $x_1 = 1$. □

使 $[x]^2$ 和 e_n 共线

NP-hard

Corollary: Minimization of convex quartic polynomial over the unit sphere is NP-hard.

Nonlinear Optimal Control: NP-hard

$x(t) = g(x,u)$, $0 \le t \le 1$

Problem: $\min_u \{ f(x(1)) : x' = g(x,u),\ 0 \le t \le 1,\ x(0) = x_0 \}$.

Consider $g(x,u) = \frac{1}{n} x \cdot \langle x, u \rangle - u$.

Lemma. Let $\|x_0\|^2 = n$. Then $\|x(t)\|^2 = n$, $0 \le t \le 1$.

Proof. Consider $\tilde{g}(x,u) = \left(\frac{xx^T}{\|x\|^2} - I \right) u$ and let $x' = \tilde{g}(x,u)$. Then

$n = \|x_0\|^2$

$$\langle x', x \rangle \quad = \quad \left\langle \left(\frac{xx^T}{\|x\|^2} - I \right) u, x \right\rangle \quad = \quad 0.$$

Thus, $\|x(t)\|^2 = \|x_0\|^2$. $= n$ Same is true for $x(t)$ defined by g. □

Note: We have enough degrees of freedom to put $x(1)$ at any position of the sphere.

Hence, our problem is: $\min\{ f(y) : \|y\|^2 = n \}$.

Descent direction for nonsmooth nonconvex function

Consider $\phi(x) = \left(1 - \frac{1}{\gamma}\right) \max\limits_{1\le i\le n} |x_i| - \min\limits_{1\le i\le n} |x_i| + |\langle a, x\rangle|,$

where $a \in Z^n_+$ and $\gamma \stackrel{\rm def}{=} \sum\limits_{i=1}^n a_i \ge 1$. Clearly, $\phi(0) = 0$.

确定下降方向

Lemma. It is NP-hard to decide if $\phi(x) < 0$ for some $x \in \mathbb{R}^n$.

Proof: 1. Assume that $\sigma \in \mathbb{R}^n$ with $\sigma_i = \pm 1$ satisfies $\langle a, \sigma\rangle = 0$. Then $\phi(\sigma) = -\frac{1}{\gamma} < 0$.

2. Assume $\phi(x) < 0$ and $\max\limits_{1\le i\le n} |x_i| = 1$. Denote $\delta = |\langle a, x\rangle|$.

Then $|x_i| > 1 - \frac{1}{\gamma} + \delta$, $i = 1, \ldots, n$.

Denoting $\sigma_i = \mathrm{sign} x_i$, we have $\sigma_i x_i > 1 - \frac{1}{\gamma} + \delta$. Therefore, $|\sigma_i - x_i| = 1 - \sigma_i x_i < \frac{1}{\gamma} - \delta$, and we conclude that

$$\begin{aligned} |\langle a, \sigma\rangle| &\le |\langle a, x\rangle| + |\langle a, \sigma - x\rangle| \le \delta + \gamma \max_{1\le i\le n} |\sigma_i - x_i| \\ &< (1-\gamma)\delta + 1 \le 1. \end{aligned}$$

Since $a \in Z^n$, this is possible iff $\langle a, \sigma\rangle = 0$. □

§2 Black-box optimization 的复杂度

2.1. Black-box optimization 的概念

Black-box optimization

Oracle: Special unit for computing function value and derivatives at test points. (0-1-2 order.) 在每个 test point x 处给出 f(x) / f(x) & ∇f(x) / f(x) & ∇f(x) & ∇²f(x) 的值

Analytic complexity: Number of calls of oracle, which is necessary (sufficient) for solving any problem from the class.

(Lower/Upper complexity bounds.)

首先需要确定 problem class!

Solution: ϵ-approximation of the minimum.

Resisting oracle: creates the worst problem instance for a particular method.

- Starts from "empty" problem.
- Answers must be compatible with the description of the problem class. 每个 test point 处 oracle 给出的 values
- The bad problem is created after the method stops. 希望 stop 得尽可能晚

upper bdd: 某个特定的 optimization scheme 的 "guaratee", 在最坏情况下至少需要多少 steps.
(valid for all problems) (若 $N \geq \cdots$, 则 $\epsilon \leq \cdots$)

lower bdd: 对于某一类型的所有的 methods, 存在一个最坏的 function, 无论选取哪种 method 都至少需要多少 steps.
即这类方法在多少步内不可能解决所有这类问题 (valid for all methods) (若 $\epsilon \leq \cdots$, 则 $N \geq \cdots$)

2.2. Lip-cont. & box-constraint 问题下 zero-order methods 的 lower bound

Bounds for Global Minimization

① 问题: Lip-cont., box-constraint ② 方法: zero-order

想找出这一类方法的 lower bound (至少多少步能达到 ϵ-accuracy)

Problem: $f^* = \min_x \{f(x) : x \in B_n\}, \quad B_n = \{x \in \mathbb{R}^n : 0 \le x \le e_n\}.$

Problem Class: $|f(x) - f(y)| \le L\|x - y\|_\infty \ \forall x, y \in B_n.$

Oracle: $f(x)$ (zero order).

Goal: Find $\bar{x} \in B_n$: $f(\bar{x}) - f^* \le \epsilon$.

Theorem: $N(\epsilon) \ge \left(\frac{L}{2\epsilon}\right)^n$.

对于 zero-order method, 这种情况是 worst case, $f(x)$ 不包含任何信息

Resisting oracle: at each test point reply $f(x) = 0$.

Assume, $N < p^n$. Then, $\exists$ ball with no questions. Hence, we can take $f^* = -\frac{L}{2p}$. Thus, $\epsilon \ge \frac{L}{2p}$. □

Corollary: Uniform Grid method is worst-case optimal.

即迭代次数 $N < p^n$ 时, answer 的精确度 $\epsilon \ge \frac{L}{2p}$

⇒ 若要精确度 $< \frac{L}{2p}$, 需要 $> p^n$ 步.

⇒ 若要精确度 $\le \epsilon$, 需要 $\ge (\frac{L}{2\epsilon})^n$ 步.

zero-order method 不能在 $(\frac{L}{2\epsilon})^n$ 步内以 ϵ 精度解决所有此类问题

2.3. Non-smooth & convex & sphere-constraint 问题下 first-order methods 的 lower bound

Nonsmooth Convex Minimization (NCM)

① 问题: Non-smooth & convex & general constraint ② 方法: first-order

想找出这一类方法的 lower bound (至少多少步能达到 ϵ-accuracy)

Problem: $f^* = \min_x \{f(x) : x \in Q\}$, where

- $Q \subseteq \mathbb{R}^n$ is a convex set: $x, y \in Q \Rightarrow [x, y] \in Q$. It is a simple set.
- $f(x)$ is sub-differentiable convex function:

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

for certain subgradient $f'(x) \in \mathbb{R}^n$.

∀x点处存在 f(x)+<f'(x), y-x>, 其在f(y)下方

在某x处可能不止一个合适的 subgradient

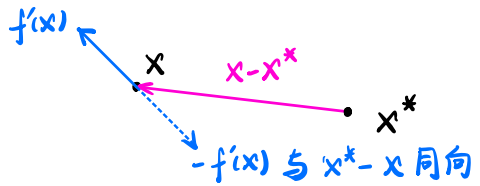
Oracle: $f(x), f'(x)$ (first order).

Solution: ϵ-approximation in function value.

(把y替换为x^*)

Main inequality: $\langle f'(x), x - x^* \rangle \geq f(x) - f^* \geq 0$, $\forall x \in Q$.

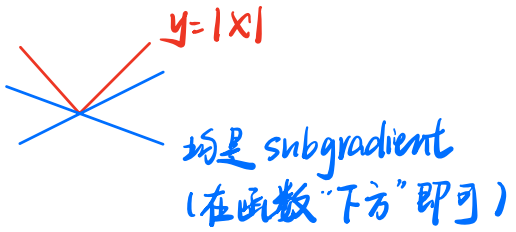
NB: Anti-subgradient decreases the distance to the optimum.



Computation of subgradients

Denote by $\partial f(x)$ the *subdifferential* of f at x.

This is the set of *all* subgradients at x.



0. If f is differentiable at x, then $\partial f(x) = \{f'(x)\}$.

1. For $f = \alpha_1 f_1 + \alpha_2 f_2$ with $\alpha_1, \alpha_2 > 0$, we have

$$\partial f(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x).$$

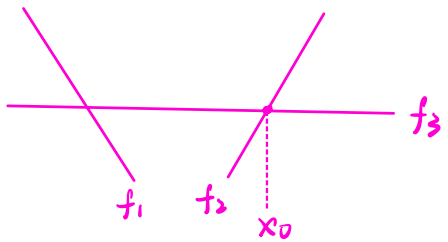
2. For $f = \max\{f_1, f_2\}$, we have

$$\partial f(x) = \text{Conv}\,\{\partial f_i(x)\}_{i \in I(x)},$$

where $I(x) = \{i : f_i(x) = f(x)\}$.

在x点处 active 的 f_i 的 $\partial f_i(x)$ 的 convex combination

$\max\{f_1, f_2, f_3\}$ 在 x_0 处 active 的为 f_2, f_3



NCM: Lower Complexity Bounds

几乎所有 first-order method 都服从

.

Let $Q \equiv \{\|x\| \le 2R\}$ and $x^{k+1} \in x^0 + \mathrm{Lin}\{f'(x^0), \dots, f'(x^k)\}$.

Consider function $f_m(x) = L \max\limits_{1 \le i \le m} x_i + \frac{\mu}{2}\|x\|^2$ with $\mu = \frac{L}{Rm^{1/2}}$.

minimize 时 $x_{m+1} \cdots x_n$ 一定取 0

$(1 \le m \le n)$

Solving the problem: $\min\limits_{\tau} \left(L\tau + \frac{\mu m}{2}\tau^2\right)$, we get

$x_1 \cdots x_m$ symmetric

$\Rightarrow$ 最优点一定相等

$\Rightarrow$ 设 $\tau_* = x_{1*} = \cdots = x_{m*}$

(二次函数)

$$\tau_* = -\frac{L}{\mu m} = -\frac{R}{m^{1/2}}, \quad f_m^* = -\frac{L^2}{2\mu m} = -\frac{LR}{m^{1/2}}, \quad \|x^*\|^2 = m\tau_*^2 = R^2.$$ (在可行域内)

$= x_{1*} = \cdots = x_{m*}$

所有 first-order method 均存在每轮迭代仅考虑一个维度的情况 (见 note P7)

NB: If $x^0 = 0$, then after k iterations we can keep $x_i = 0$ for $i > k$.

general

~~Lipschitz continuity~~ case: $f_{k+1}(x^k) - f_{k+1}^* \ge -f_{k+1}^* = \frac{LR}{(k+1)^{1/2}}$.

取 $m = k+1$

$= L \cdot 0 + \frac{\mu}{2}\|x^k\| \ge 0$

Strong convexity case: $f_{k+1}(x^k) - f_{k+1}^* \ge -f_{k+1}^* = \frac{L^2}{2(k+1)\cdot\mu}$.

Both lower bounds are *exact*!

$f(y) \ge f(x) + \langle f'(x), y - x\rangle + \frac{1}{2}\mu\|y - x\|^2$

$\mu > 0$, $y \in Q$

即迭代次数 $N < k$ 时，answer 的精确度 $\varepsilon \ge \frac{LR}{(k+1)^{1/2}}$

$\Rightarrow$ 若要精确度 $< \frac{LR}{(k+1)^{1/2}}$，需要 $> k$ 步.

$\Rightarrow$ 若要精确度 $\le \varepsilon$，需要 $\ge (\frac{LR}{\varepsilon})^2 - 1$ 步. (此时不用管 m 取多少)

first-order method 不能在 $(\frac{LR}{\varepsilon})^2 - 1$ 步内以 ε 精度解决所有此类问题

2.4. Non-smooth & convex & functional-constraint 问题下 Subgradient method 的 upper bound

Subgradient Method (SG)

Problem: $\min_{x \in Q} \{f(x) : g(x) \le 0\}$,

where Q is a closed convex set, and convex $f, g \in C_L^{0,0}(Q)$.

含 h 为 step size　→ $g(x^k) \le 0$ 不满足　算子: projection to Q　向 $g(x^k)$ 负梯度方向走 $\frac{g(x^k)}{\|g'(x^k)\|}$

SG: If $\frac{g(x^k)}{\|g'(x^k)\|} > h$ then **a)** $x^{k+1} = \pi_Q\left(x^k - \frac{g(x^k)}{\|g'(x^k)\|^2} g'(x^k)\right)$,

试图 minimize $g(x)$

else **b)** $x^{k+1} = \pi_Q\left(x^k - \frac{h}{\|f'(x^k)\|} f'(x^k)\right)$.

向 $f(x^k)$ 负梯度方向走 h

Denote $f_N^* = \min_{0 \le k \le N} \{f(x^k) : k \in \mathbf{b})\}$. Let $N = N_a + N_b$.

选择该 step size 以保证误差为 ϵ

Theorem: If $N > \frac{1}{h^2}\|x^0 - x^*\|^2$, then $f_N^* - f^* \le hL$. $(h = \frac{\epsilon}{L}.)$

Proof: Denote $r_k = \|x^k - x^*\|$.

a): $r_{k+1}^2 - r_k^2 \le -\frac{2g(x^k)}{\|g'(x^k)\|^2}\langle g'(x^k), x^k - x^*\rangle + \frac{g^2(x^k)}{\|g'(x^k)\|^2} \le -h^2$.

b): $r_{k+1}^2 - r_k^2 \le -\frac{2h\langle f'(x^k), x^k - x^*\rangle}{\|f'(x^k)\|} + h^2 \le -\frac{2h}{L}(f(x^k) - f^*) + h^2$.

Thus, $N_b \frac{2h}{L}(f_N^* - f^*) \le r_0^2 + h^2(N_b - N_a) = r_0^2 + h^2(2N_b - N)$. □

证明了 ① 前一页的 LB 为 exact ② 对于这类问题, subgradient method 为 optimal

2.5. Smooth & convex 问题下 first-order methods 的 lower bound

Smooth Convex Minimization (SCM)

Lipschitz continuous gradient

Lipschitz-continuous gradient: $\|f'(x) - f'(y)\| \leq L\|x - y\|$.

Geometric interpretation: for all $x, y \in \operatorname{dom} F$ we have

$$\begin{aligned} 0 &\leq f(y) - f(x) - \langle f'(x), y - x\rangle \quad \text{(convexity)} \\ &= \int_0^1 \langle f'(x + t(y-x)) - f'(x), y - x\rangle dt \leq \tfrac{L}{2}\|x - y\|^2. \end{aligned}$$

$= \int_0^1 \langle f'(x+t(y-x)), y-x\rangle dt$

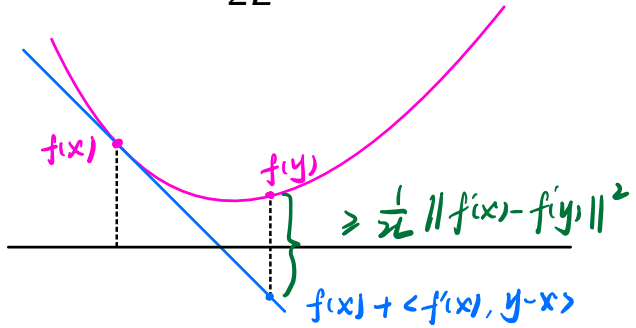
$\leq tL \cdot (y-x)$

Sufficient condition: $0 \preceq f''(x) \preceq L \cdot I_n$, $x \in \operatorname{dom} f$.

Equivalent definition:

$$f(y) \geq f(x) + \langle f'(x), y - x\rangle + \tfrac{1}{2L}\|f'(x) - f'(y)\|^2.$$

Hint: Prove first that $f(x) - f^* \geq \frac{1}{2L}\|f'(x)\|^2$.



SCM: Lower complexity bounds

Consider the family of functions ($k \leq n$):

$$f_k(x) = \frac{1}{2}\left[x_1^2 + \sum_{i=1}^{k-1}(x_i - x_{i+1})^2 + x_k^2\right] - x_1 \equiv \frac{1}{2}\langle A_k x, x\rangle - x_1.$$

$f_k'(x) = A_k x - e_1$

Let $\mathbb{R}_k^n = \{x \in \mathbb{R}^n : x_i = 0,\ i > k\}$. Then $f_{k+p}(x) = f_k(x)$, $x \in \mathbb{R}_k^n$.

Clearly, $0 \leq \underline{\langle A_k h, h\rangle} \leq h_1^2 + \sum_{i=1}^{k-1} 2\,(h_i^2 + h_{i+1}^2) + h_k^2 \leq 4\|h\|^2$,

$= \|f'(x) - f'(y)\|^2,\ h = x - y$

$$A_k = \begin{pmatrix} \begin{matrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{matrix} & & \begin{matrix} & 0 & \\ & & \\ & & \end{matrix} \\ \begin{matrix} & \cdots & \\ & & \\ & 0 & \end{matrix} & & \begin{matrix} & \cdots & \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{matrix} \end{pmatrix} \left.\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}\right\} \begin{matrix} k \text{ lines} \\ \\ \\ \\ \\ \end{matrix}$$

$$\begin{pmatrix} 0_{n-k,k} & 0_{n-k,n-k} \end{pmatrix},$$

Hence, $A_k x = e_1$ has the solution $\bar{x}_i^k = \begin{cases} \frac{k+1-i}{k+1}, & 1 \le i \le k, \\ 0, & i > k. \end{cases}$.

Thus $f_k^* = \frac{1}{2}\langle A_k \bar{x}^k, \bar{x}^k \rangle - \langle e_1, \bar{x}^k \rangle = -\frac{1}{2}\langle e_1, \bar{x}^k \rangle = -\frac{k}{2(k+1)}$, and

$$\| \bar{x}^k \|^2 = \sum_{i=1}^{k} \left(\frac{k+1-i}{k+1} \right)^2 = \frac{1}{(k+1)^2} \sum_{i=1}^{k} i^2 = \frac{k(2k+1)}{6(k+1)}.$$

Let $x^0 = 0$ and $p \le n$ is fixed.

Lemma. If $x^k \in \mathcal{L}_k \stackrel{\text{def}}{=} \text{Lin}\{f_p'(x^0), \ldots, f_p'(x^{k-1})\}$, then $\mathcal{L}_k \subseteq \mathbb{R}_k^n$.

Proof: $x^0 = 0 \in \mathbb{R}_0^n, f_p'(0) = -e_1 \in \mathbb{R}_1^n \Rightarrow x^1 \in \mathbb{R}_1^n, f_p'(x_1) \in \mathbb{R}_2^n, etc. \square$

Corollary 1: $f_p(x^k) = f_k(x^k) \ge f_k^*$.

⇒ 若要精确度 $< \frac{1}{O(k^2)}$，需要 $> k$ 步.
⇒ 若要精确度 $\le \varepsilon$，需要 $\ge \frac{1}{O(\varepsilon^{\frac{1}{2}})}$ 步.
↖ convergence rate 为 $O(\frac{1}{k^2})$

Corollary 2: Take $p = 2k+1$. Then

$$\frac{f_p(x^k) - f_p^*}{L\|x^0 - \bar{x}^p\|^2} \ge \frac{1}{4}\left[-\frac{k}{2(k+1)} + \frac{2k+1}{2(2k+2)} \right] / \left[\frac{(2k+1)(4k+3)}{12(k+1)} \right] = \boxed{\frac{3}{4(2k+1)(4k+3)}}.$$

$$\| x^k - \bar{x}^p \|^2 \ge \sum_{i=k+1}^{2k+1} (\bar{x}_i^{2k+1})^2 = \frac{(2k+3)(k+2)}{24(k+1)} \boxed{\ge \frac{1}{8}\|\bar{x}^p\|^2}.$$ 但不能确保 x^k 收敛至 x^*

Some remarks

1. The rate of convergence of *any* Black-Box gradient methods as applied to $f \in C^{1,1}$ cannon be higher than $O(\frac{1}{k^2})$.

2. We cannot guarantee *any* rate of convergence in the argument.

3. Let $A = LL^T$ and $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$. Then

$$f(x) - f^* = \tfrac{1}{2}\|L^T x - d\|^2, \text{ where } d = L^T x^*.$$

Thus, the residual of the linear system $L^T x = b$ cannot be decreased faster than with the rate $O(\frac{1}{k})$

(provided that we are allowed to multiply by L and L^T.)

4. Optimization problems with nontrivial linear *equality constraints* cannot be solved faster than with the rate $O(\frac{1}{k})$.

2.6. Smooth & convex 问题 first-order methods: PGM, DGM, 和两者结合 (estimating sequence)

Smooth Minimization with Simple Constraints

Intuition: 每次用一个 quadratic function 近似, 最小化该 function, 其 minimum 是 $f(T_M)$ 的 upper bound

Consider the problem: $\min_x \{f(x) : x \in Q\}$,

where convex $f \in C_L^{1,1}(Q)$, and Q is a simple closed convex set (allows projections).

Gradient mapping: for $M > 0$ define

$$T_M(x) = \arg\min_{y \in Q} [f(x) + \langle f'(x), y - x \rangle + \frac{M}{2}\|x - y\|^2].$$

regularization

(*)

If $M \geq L$, then

$$f(T_M(x)) \leq f(x) + \langle f'(x), T_M(x) - x \rangle + \frac{M}{2}\|x - T_M(x)\|^2]. \quad (\#)$$

descent lemma

Reduced gradient: $g_M(x) = M \cdot (x - T_M(x))$.

Since $\langle f'(x) + M(T_M(x) - x), y - T_M(x) \rangle \geq 0$ for all $y \in Q$, (*) 的 FOC ($\frac{\partial}{\partial y}(*)|_{y=T_M(x)}$)

$$f(x) - f(T_M(x)) \geq \frac{M}{2}\|x - T_M(x)\|^2 = \frac{1}{2M}\|g_M(x)\|^2, \quad (\to 0)$$

$$f(y) \geq f(x) + \langle f'(x), T_M(x) - x \rangle + \langle f'(x), y - T_M(x) \rangle$$

$$\geq f(T_M(x)) - \frac{1}{2M}\|g_M(x)\|^2 + \langle g_M(x), y - T_M(x) \rangle.$$

$\geq \langle f(x), x - T_M(x) \rangle - \frac{M}{2}\|x - T_M(x)\|^2$

Primal Gradient Method (PGM)

Main scheme: $x^0 \in Q$, $x^{k+1} = T_L(x^k)$, $k \geq 0$.

Primal interpretation: $x^{k+1} = \pi_Q\left(x^k - \frac{1}{L} f'(x^k)\right)$.

Rate of convergence. $f(x^k) - f(x^{k+1}) \geq \frac{1}{2L}\|g_L(x^k)\|^2$.

$$\begin{aligned} f(T_L(x)) - f^* &\leq \frac{1}{2L}\|g_L(x)\|^2 + \langle g_L(x), T_L(x) - x^* \rangle \\ &\leq \frac{1}{2L}(\|g_L(x)\| + LR)^2 - \frac{L}{2}R^2. \end{aligned}$$

Hence, $\|g_L(x)\| \geq \left[2L(f(T_L(x)) - f^*) + L^2R^2\right]^{1/2} - LR$

$$= \frac{2L(f(T_L(x)) - f^*)}{[2L(f(T_L(x)) - f^*) + L^2R^2]^{1/2} + LR} \geq \frac{c}{R} \cdot (f(T_L(x)) - f^*).$$

Thus, $f(x^k) - f(x^{k+1}) \geq \frac{c^2}{LR^2}(f(x^{k+1}) - f^*)^2$.

Similar situation: $a'(t) = -a^2(t) \Rightarrow a(t) \approx \frac{1}{t}$.

Conclusion: PGM converges as $O(\frac{1}{k})$. This is far from the lower complexity bounds.

Dual Gradient Method (DGM)

Model: Let $\lambda_i^k \geq 0$, $i = 0, \ldots, k$, and $S_k \stackrel{\text{def}}{=} \sum_{i=0}^{k} \lambda_i^k$. Then

$$S_k f(y) \geq \mathcal{L}_{\lambda^k}(y) \stackrel{\text{def}}{=} \sum_{i=0}^{k} \lambda_i^k [f(x^i) + \langle f'(x^i), y - x^i \rangle], \quad y \in Q.$$

DGM: $\boxed{x^{k+1} = \arg\min_{y \in Q} \left\{ \psi_k(y) \stackrel{\text{def}}{=} \mathcal{L}_{\lambda^k}(y) + \frac{M}{2}\|y - x^0\|^2 \right\}.}$

Let us choose $\lambda_i^k \equiv 1$ and $M = L$. We prove by induction

$$(*): \quad F_k^* \stackrel{\text{def}}{=} \sum_{i=0}^{k} f(x^{i+1}) \leq \psi_k^* \stackrel{\text{def}}{=} \min_{y \in Q} \psi_k(y). \quad (\leq (k+1)f^* + \tfrac{L}{2}R^2)$$

1. $k = 0$. Then $(*)$ is true.
2. Assume $(*)$ is true for some $k \geq 0$. Then

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

$$\geq \psi_k^* + \tfrac{L}{2}\|x^{k+1} - x^k\|^2 + f(x^k) + \langle f'(x^k), x^{k+1} - x^k \rangle \geq \psi_k^* + f(x^{k+1}).$$

Thus, $\frac{1}{k+1} \sum_{i=0}^{k} f(x^{i+1}) \leq f^* + \boxed{\frac{LR^2}{2(k+1)}}$.

Some remarks

1. Dual gradient method works with the *model* of the objective function.

2. Both primal and dual methods have the same rate of convergence $O(\frac{1}{k})$. It is not optimal.

May be we can combine them in order to get a better rate?

Comparing PGM and DGM

Primal Gradient method

- Monotonically improves the current state using the local model of the objective.
- **Interpretation:** Practitioners, industry.

Dual Gradient Method

- The main goal is to construct a model of the objective.
- It is updated by a new experience collected around the predicted test points (x_k).
- Practical verification of the advices is not essential for the procedure.
- **Interpretation:** Science.

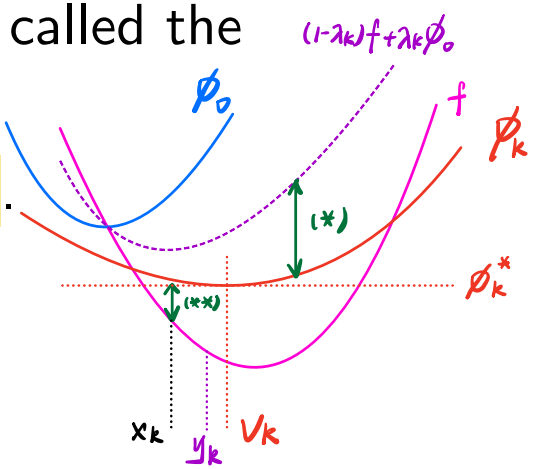
Hint: Combination of theory and practice should give better results.

Estimating sequences

λ_k：初始model在第k个迭代中的占比
(每轮后为α_k的比例下降)
$\alpha_k^2 = (1-\alpha_k)\lambda_k$

Def. Sequences $\{\phi_k(x)\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$, $\lambda_k \geq 0$, are called the *estimating sequences* if $\lambda_k \to 0$ and $\forall x \in Q$, $k \geq 0$,

$$(*): \quad \phi_k(x) \leq (1-\lambda_k) f(x) + \lambda_k \phi_0(x).$$



Lemma: If $(**): f(x^k) \leq \phi_k^* \equiv \min\limits_{x \in Q} \phi_k(x)$, then

$$f(x^k) - f^* \leq \lambda_k [\phi_0(x^*) - f^*] \to 0.$$

Proof. $f(x^k) \leq \phi_k^* = \min\limits_{x \in Q} \phi_k(x) \leq \min\limits_{x \in Q} [(1-\lambda_k) f(x) + \lambda_k \phi_0(x)]$

$$\leq (1-\lambda_k) f(x^*) + \lambda_k \phi_0(x^*). \quad \square$$

Rate of $\lambda_k \to 0$ defines the rate of $f(x^k) \to f^*$.

Questions

- How to construct the estimating sequences?
- How we can ensure (**)?

Updating the estimating sequences

Lemma: Let $\phi_0(x) = \frac{L}{2}\|x - x^0\|^2$, $\lambda_0 = 1$, $\{y^k\}_{k=0}^{\infty}$ a sequence in Q, and $\{\alpha_k\}_{k=0}^{\infty}$: $\alpha_k \in (0,1)$, $\sum\limits_{k=0}^{\infty} \alpha_k = \infty$. Then $\{\phi_k(x)\}_{k=0}^{\infty}$, $\{\lambda_k\}_{k=0}^{\infty}$:

$$\lambda_{k+1} = (1 - \alpha_k)\lambda_k,$$

$$\phi_{k+1}(x) = (1 - \alpha_k)\phi_k(x) + \alpha_k[f(y^k) + \langle f'(y^k), x - y^k\rangle]$$

are estimating sequences.

Proof: $\phi_0(x) \le (1 - \lambda_0)f(x) + \lambda_0\phi_0(x) \equiv \phi_0(x)$.

If (*) holds for some $k \ge 0$, then

$$\begin{aligned}
\phi_{k+1}(x) \le & \quad (1 - \alpha_k)\phi_k(x) + \alpha_k f(x) \\
= & \quad (1 - (1 - \alpha_k)\lambda_k)f(x) + (1 - \alpha_k)(\phi_k(x) - (1 - \lambda_k)f(x)) \\
\le & \quad (1 - (1 - \alpha_k)\lambda_k)f(x) + (1 - \alpha_k)\lambda_k\phi_0(x) \\
= & \quad (1 - \lambda_{k+1})f(x) + \lambda_{k+1}\phi_0(x). \quad \square
\end{aligned}$$

Updating the points

Denote $\phi_k^* = \min\limits_{x \in Q} \phi_k(x)$, $v^k = \arg\min\limits_{x \in Q} \phi_k(x)$. Suppose $\phi_k^* \geq f(x^k)$.

Then

$$\phi_{k+1}^* = \min_{x \in Q} \left\{ (1 - \alpha_k)\phi_k(x) + \alpha_k [f(y^k) + \langle f'(y^k), x - y^k \rangle] \right\} \geq$$

$$\min_{x \in Q} \left\{ (1 - \alpha_k)[\phi_k^* + \tfrac{\lambda_k L}{2} \|x - v_k\|^2] + \alpha_k [f(y^k) + \langle f'(y^k), x - y^k \rangle] \right\}$$

$$\geq \min_{x \in Q} \{ f(y^k) + \tfrac{(1-\alpha_k)\lambda_k L}{2} \|x - v_k\|^2$$

$$+ \langle f'(y^k), \alpha_k(x - y^k) + (1 - \alpha_k)(x^k - y^k) \rangle \}$$

$$\left(y_k \overset{\text{def}}{=} (1 - \alpha_k)x^k + \alpha_k v^k = x^k + \alpha_k(v^k - x^k) \right)$$

$$= \min_{x \in Q} \{ f(y^k) + \tfrac{(1-\alpha_k)\lambda_k L}{2} \|x - v_k\|^2 + \alpha_k \langle f'(y^k), x - v^k \rangle \}$$

$$= \min_{\substack{y = x^k + \alpha_k(x - x^k) \\ x \in Q}} \{ f(y^k) + \tfrac{(1-\alpha_k)\lambda_k L}{2\alpha_k^2} \|y - y_k\|^2 + \langle f'(y^k), y - y^k \rangle \} \overset{(?)}{\geq} f(x^{k+1})$$

Answer: $\alpha_k^2 = (1 - \alpha_k)\lambda_k$, $x_{k+1} = T_L(y_k)$.

Optimal method

Choose $v^0 = x^0 \in Q$, $\lambda_0 = 1$, $\phi_0(x) = \frac{L}{2}\|x - x^0\|^2$.

For $k \geq 0$ iterate:

- Compute α_k : $\alpha_k^2 = (1 - \alpha_k)\lambda_k \equiv \lambda_{k+1}$. $:= (1-\alpha_k)\lambda_k$
- Define $y_k = (1 - \alpha_k)x^k + \alpha_k v^k$.
- Compute $x^{k+1} = T_L(y^k)$.
- $\phi_{k+1}(x) = (1 - \alpha_k)\phi_k(x) + \alpha_k[f(y^k) + \langle f'(y^k), x - y^k\rangle]$.

Convergence: Denote $a_k = \lambda_k^{-1/2}$. Then

$$a_{k+1} - a_k = \frac{\lambda_k^{1/2} - \lambda_{k+1}^{1/2}}{\lambda_k^{1/2}\lambda_{k+1}^{1/2}} = \frac{\lambda_k - \lambda_{k+1}}{\lambda_k^{1/2}\lambda_{k+1}^{1/2}(\lambda_k^{1/2} + \lambda_{k+1}^{1/2})} \geq \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k\lambda_{k+1}^{1/2}} = \frac{\alpha_k}{2\lambda_{k+1}^{1/2}} = \frac{1}{2}.$$

Thus, $a_k \geq 1 + \frac{k}{2}$. Hence, $\lambda_k \leq \frac{4}{(k+2)^2}$.

Interpretation

1. $\phi_k(x)$ accumulates all previously computed information about the objective. This is a current *model* of our problem.

2. $v^k = \arg\min_{x \in Q} \phi_k(x)$ is a prediction of the optimal strategy.

3. $\phi_k^* = \phi_k(v^k)$ is an estimate of the optimal value.

4. Acceleration condition: $f(x^k) \leq \phi_k^*$. We need a firm, which is at least as good as the best theoretical prediction.

5. Then we create a startup $y^k = (1 - \alpha_k)x^k + \alpha_k v^k$, and allow it to work one year.

6. Theorem: For the next year, its performance will be at least as good as the new theoretical prediction. And we can continue!

Acceleration result: 10 years instead 100.

Who is in a right position to arrange **5**? Government, political institutions.