Lecture 1: Intrinsic complexity of Black-Box Optimization

Yurii Nesterov, CORE/INMA (UCL)

September 6, 2012

Outline

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

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

Another variant: $\sum\limits_{i=2}^{n} a_i x_i = a_1$.

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

Immediate consequence: quartic polynomial

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

Proof: Consider the following function:

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

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 $[x]_i^2 = x_i^2$, $i = 1, \ldots, n$.

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

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

Nonlinear Optimal Control: NP-hard

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

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

Thus, $\|x(t)\|^2 = \|x_0\|^2$. Same is true for $x(t)$ defined by g. $\square$
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 of 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 $\gamma \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} a_i \ \ge 1$. Clearly, $\phi(0) = 0$.

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

Proof: 1. Assume that $\sigma \in R^n$ 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 $|x_i| > 1 - \frac{1}{\gamma} + \delta$, $i = 1, \dots, n$.

Denoting $\sigma_i = \text{sign} x_i$, 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 $\langle a, \sigma \rangle = 0$.

Black-box optimization

Oracle: Special unit for computing function value and derivatives at test points. (0-1-2 order.)

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

(Lower/Upper complexity bounds.)

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.
- The bad problem is created after the method stops.

Bounds for Global Minimization

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

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

Proof. Divide B_n on p^n l_∞-balls of radius $\frac{1}{2p}$.

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

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

Corollary: Uniform Grid method is worst-case optimal.

Nonsmooth Convex Minimization (NCM)

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

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

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

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

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

Solution: ϵ-approximation in function value.

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.

NCM: Lower Complexity Bounds

.

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

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

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

$$\tau_* = -\frac{L}{\mu m} = -\frac{R}{m^{1/2}}, \; f_m^* = -\frac{L^2}{2\mu m} = -\frac{LR}{m^{1/2}}, \; \|x^*\|^2 = m\tau_*^2 = R^2.$$

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

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

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

Both lower bounds are <u>*exact*</u>!

Subgradient Method

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

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

Method If $\frac{g(x^k)}{\|g'(x^k)\|} > h$ then **a)**

$x^{k+1} = \pi_Q \left(x^k - \frac{g(x^k)}{\|g'(x^k)\|^2} g'(x^k) \right)$,

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

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

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

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

Smooth Convex Minimization (SCM)

Lipschitz-continuous gradient: $\|f'(x) - f'(y)\| \leq L\|x - y\|$.
Geometric interpretation: for all $x, y \in \operatorname{dom} F$ we have

$$\begin{array}{rcl} 0 & \leq & f(y) - f(x) - \langle f'(x), y - x \rangle \\ & = & \int\limits_0^1 \langle f'(x + \tau(y - x) - f'(x), y - x \rangle dt \leq \frac{L}{2}\|x - y\|^2. \end{array}$$

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

Equivalent definition:

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

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

SCM: Lower complexity bounds

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

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

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

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

$$A_k = \left(\begin{array}{cc} \left. \begin{array}{cccccc} 2 & -1 & 0 & & & \\ -1 & 2 & -1 & & 0 & \\ 0 & -1 & 2 & & & \\ & \dots & & & \dots & \\ & 0 & & -1 & 2 & -1 \\ & & & 0 & -1 & 2 \end{array} \right\} & k \text{ lines} \\ 0_{n-k,k} & 0_{n-k,n-k} \end{array} \right),$$

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

Proof: $x^0 = 0 \in R_0^n, f_p'(0) = -e_1 \in R_1^n \Rightarrow x^1 \in R_1^n, f_p'(x_1) \in R_2^n, \Box$

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

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

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

$$\| x^k - \bar{x}^p \|^2 \ge \sum_{i=k+1}^{2k+1} (\bar{x}_i^{2k+1})^2 = \frac{(2k+3)(k+2)}{24(k+1)} \ge \frac{1}{8}\|\bar{x}^p\|^2.$$

Some remarks

1. The rate of convergence of *any* Black-Box gradient methods as applied to $f \in C^{1,1}$ cannon be high 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

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

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 $O(\frac{1}{k})$.

Methods for Smooth Minimization with Simple Constraints

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

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

If $M \geq L$, then

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

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 $y \in Q$,

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

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

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

Primal Gradient Method (PGM)

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

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

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

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

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 $\lambda_i^k \geq 0$, $i = 0, \dots, k$, and $S_k \stackrel{\mathrm{def}}{=} \sum\limits_{i=0}^{k} \lambda_i^k$. Then

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

Our method:

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

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

$(*):\quad F_k^* \stackrel{\mathrm{def}}{=} \sum\limits_{i=0}^{k} f(y^i) \leq \psi_k^* \stackrel{\mathrm{def}}{=} \min\limits_{y \in Q} \psi_k(y). \quad (\leq (k+1)f^* + \frac{L}{2}R^2)$

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

$$\begin{aligned}\psi_{k+1}^* &= \min_{y \in Q} \left[\psi_k(y) + f(x^k) + \langle f'(x^k), y - x^k \rangle \right] \\ &\geq \min_{y \in Q} \left[\psi_k^* + \frac{L}{2}\|y - x^k\|^2 + f(x^k) + \langle f'(x^k), y - x^k \rangle \right].\end{aligned}$$

We can take $y^{k+1} = T_L(x_k)$. Thus, $\frac{1}{k+1} \sum\limits_{i=0}^{k} f(y^i) \leq f^* + \frac{LR^2}{2(k+1)}$.

Some remarks

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

2. The minimizing sequence $\{y^k\}$ is not necessary for the algorithmic scheme. We can generate it if necessary.

3. Both primal and dual method 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 (y_k) is not essential for the procedure.
- **Interpretation:** Science.

Hint: Combination of theory and practice should give better results

Estimating sequences

Def. A sequences $\{\phi_k(x)\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$, $\lambda_k \geq 0$ are called the <u>*estimating sequences*</u> 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 $(**) : f(x^k) \leq \phi_k^* \equiv \min_{x \in Q} \phi_k(x)$, then

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

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

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

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 estimating sequences

Let $\phi_0(x) = \frac{L}{2}\|x - x^0\|^2$, $\lambda_0 = 1$, $\{y^k\}_{k=0}^{\infty}$ is a sequence in Q, and $\{\alpha_k\}_{k=0}^{\infty}$: $\alpha_k \in (0,1)$, $\sum_{k=0}^{\infty} \alpha_k = \infty$. 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: $\phi_0(x) \leq (1-\lambda_0)f(x) + \lambda_0\phi_0(x) \equiv \phi_0(x)$.

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

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

Updating the points

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

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

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

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

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

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

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

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

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 : $\alpha_k^2 = (1-\alpha_k)\lambda_k \equiv \lambda_{k+1}$.
- 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: 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.