New perspectives for increasing efficiency of optimization schemes

Yurii Nesterov, CORE/INMA (UCL)

January 7, 2016 (Big Data Conference, UCL, London)

Joint results with S. Stich (ETH Zurich)

Thirty years ago ...

Problem: $\min\limits_{x \in \mathbb{R}^n} f(x)$, f is convex and

$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, $x, y \in \mathbb{R}^n$.

Fast gradient method (N.1984): $x_0 \in \mathbb{R}^n$,

$y_k = x_k + \frac{k}{k+2}(x_k - x_{k-1})$, $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$, $k \geq 0$.

Result: $f(x_k) - f^* \leq \frac{2LR^2}{k(k+1)}$, $k \geq 1$. (Optimal)

Compare: Heavy ball method (B.Polyak, 1964)

$x_{k+1} = x_k + \alpha_k(x_k - x_{k-1}) - \beta_k \nabla f(x_k)$, $k \geq 0$.
(Convergence analysis for QP)

Applications (1983-2003): Nothing ...

Twenty years later ...

Problem: $\min_{x \in Q} f(x)$, f is convex, Q is convex, and

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

Prox-function: strongly convex $d(x)$, $x \in Q$.

Fast gradient method (N.03): $v_0 = x_0 \in \mathbb{R}^n$,

$$v_k = \arg\min_{x \in Q} \left\{ d(x) + \sum_{i=0}^{k-1} \frac{i+1}{2L}[f(y_i) + \langle \nabla f(y_i), x - y_i \rangle] \right\},$$

$$y_k = \frac{k}{k+3} v_k + \frac{2}{k+3} x_k, \quad x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k), \; k \geq 0.$$

Result: $f(x_k) - f^* \leq \frac{2LR^2}{k(k+1)}$, $k \geq 1$.

Applications (2003 - 2015(?)): Smooth approximations of nonsmooth functions.

Smoothing technique

Problem: $\min_{x \in Q} f(x)$, $\operatorname{diam} Q = D_1$.

Model: $f(x) = \max_{u \in U} \{\langle Ax, u\rangle - \phi(u)\}$.

Smoothing: $f_\mu(x) = \max_{u \in U} \{\langle Ax, u\rangle - \phi(u) - \mu d_2(u)\}$,
where d_2 is strongly convex on U, $\operatorname{diam} U = D_2$.

Then $\|\nabla f_\mu(x) - \nabla f_\mu(y)\|_* \leq \frac{1}{\mu}\|A\|^2 \cdot \|x - y\|$, $x, y \in \mathbb{R}^n$, $\mu > 0$,

where $\|A\| = \max_{x,u} \{\langle Ax, u\rangle : \ \|x\| \leq 1, \ \|u\| \leq 1\}$.

Complexity: Choose $\mu = O(\epsilon)$.

Then we get ϵ-solution in $O(\frac{1}{\epsilon}\|A\|D_1 D_2)$ iterations.

NB: Subgradient schemes need $O(\frac{1}{\epsilon^2}\|A\|^2 D_1^2 D_2^2)$ iteratons.

Ten years later ...

Huge-scale problems $\Rightarrow$ *Coordinate descent methods*

Problem: $\min\limits_{x \in \mathbb{R}^n} f(x)$, with objective satisfying conditions

$$|\nabla_i f(x + h e_i) - \nabla_i f(x)| \leq L_i |h|, \; \forall x, h \in \mathbb{R}^n. \quad (\textbf{NB:}\; L_i \leq L_{\nabla f}.)$$

Hence: $f(x) - f\left(x - \frac{1}{L_i} \nabla_i f(x) e_i\right) \geq \frac{1}{2L_i} \left(\nabla_i f(x)\right)^2, \; i = 1 : n.$

Consequences [N.12]: Denote $S_\alpha = \sum\limits_{i=1}^{n} L_i^\alpha, \; \alpha \in [0,1]$.

- Choose i with probability $\pi_i = \frac{1}{S_1} L_i$. Then
  $\mathcal{E}\left(f(x_+)\right) \leq f(x) - \frac{1}{2S_1} \|\nabla f(x)\|_{[0]}^2$
  $\Rightarrow \mathcal{E}\left(f(x_k)\right) - f^* \leq \frac{S_1 R_0^2}{k}$.
  $$\left(\|x\|_{[\alpha]}^2 = \sum_{i=1}^{n} L_i^\alpha \left(x^{(i)}\right)^2, \quad \|g\|_{[\alpha]}^2 = \sum_{i=1}^{n} L_i^{-\alpha} \left(g^{(i)}\right)^2. \right)$$
- Choose i with probability $\pi_i = \frac{1}{n}$. Then

Fast Coordinate Descent

Random generator: For $\beta \in [0,1]$, get $j = \mathcal{R}_\beta(L) \in \{1:n\}$ with probabilities $\pi_\beta[i] \equiv \mathrm{Prob}(j=i) \stackrel{\mathrm{def}}{=} \frac{1}{S_\beta} L_i^\beta, \quad i \in \{1:n\}$.

- [N.12] $\beta = 0$: $\mathcal{E}(f(x_k)) - f^* \leq 2(\frac{n}{k+1})^2 L_{\max} R_{[0]}^2$.
- [Lee, Sidford 13] $\beta = 1$: $\mathcal{E}(f(x_k)) - f^* \leq 2\frac{nS_1}{(k+1)^2} R_{[0]}^2$.
- [N., Stich 15] $\beta = \frac{1}{2}$: $\mathcal{E}(f(x_k)) - f^* \leq 2\left(\frac{S_{1/2}}{k+1}\right)^2 R_{[0]}^2$.

Fast CD: Choose $v_0 = x_0 \in \mathbb{R}^n$. Set $A_0 = 0$

1) Choose active coordinate $i_t = \mathcal{R}_{1/2}(L)$.

2) Solve $a_{t+1}^2 S_{1/2}^2 = A_t + a_{t+1}$. Set $A_{t+1} = A_t + a_{t+1}$, $\alpha_t = \frac{a_{t+1}}{A_{t+1}}$.

3) Set $y_t = (1-\alpha_t)x_t + \alpha_t v_t$, $x_{t+1} = y_t - \frac{1}{L_{i_t}} \nabla_{i_t} f(y_t) e_{i_t}$,

$$v_{t+1} = v_t - \frac{a_{t+1} S_{1/2}}{L_{i_t}^{1/2}} \nabla_{i_t} f(y_t) e_{i_t}.$$

NB: Each iteration needs $O(n)$ a.o. $\Rightarrow$ Not for Huge Scale.

Complexity

For getting ϵ-accuracy we need $\frac{S_{1/2} R_{[0]}}{\epsilon^{1/2}}$ iterations $\leq \frac{n L_{\nabla f}^{1/2} R_{[0]}}{\epsilon^{1/2}}$.

Main question: When CD-oracle is n times cheaper?

NB: For FGM, complexity oracle/method is often <u>unbalanced</u>.

Model: $f(x) = F(Ax, x)$, where $F(s,x) : \mathbb{R}^{m+n} \to \mathbb{R}$.

Main assumption: $F(s,x)$ can be computed in $O(m+n)$ a.o.
(And $\nabla F \in \mathbb{R}^{m+n}$ too!)

Consequences: Let $A = (A_1, \dots, A_n)$.

- After coordinate move, product Ax can be computed in $O(m)$ a.o.
- If Ax is known, $\nabla_i f(x) = \langle A_i, \nabla_s F \rangle + \nabla_{x_i} F$ can be computed in $O(m)$ a.o.
- If Ax and Ay are known, $\alpha Ax + \beta Ay$ can be computed in

Example 1: Unconstrained quadratic minimization

Let $A = A^T \succ 0 \in \mathbb{R}^{n\times n}$ is dense.

Define $F(s,x) = \frac{1}{2}\langle s,x\rangle - \langle b,x\rangle$.

Then $f(x) = \frac{1}{2}\langle Ax,x\rangle - \langle b,x\rangle$. We have

$$T_{CD} = O\left(\frac{nS_{1/2}}{\epsilon^{1/2}}R_{[0]}\right) \quad \leq \quad T_{FGM} = O\left(\frac{n^2\lambda_{\max}^{1/2}(A)}{\epsilon^{1/2}}R_{[0]}\right).$$

NB: We use inequality $L_i^{1/2} \leq \lambda_{\max}^{1/2}(A)$, $i \in \{1:n\}$.

For some cases it is too weak.

Example: $0 < \gamma_1 \leq A^{(i,j)} \leq \gamma_2$, $i,j \in \{1:n\}$.

Then $L_i^{1/2} \leq \gamma_2^{1/2}$ and $\lambda_{\max}(A) \geq \gamma_1\lambda_{\max}(1_n 1_n^T) = n\gamma_2$.

We gain $O(n^{1/2})$ in the total number of operations.

Example 2: Smoothing technique

Consider function $f(x) = \max\limits_{u \in Q} \{ \langle Ax, u \rangle - \phi(u) \}$,

where $Q \subset \mathbb{R}^m$ is closed convex and bounded.

Define $f_\mu(x) = \max\limits_{u \in Q} \{ \langle Ax, u \rangle - \phi(u) - \mu d(u) \}$, $\mu > 0$.

Main assumption: $F(s) = \max\limits_{u \in Q} \{ \langle s, u \rangle - \phi(u) - \mu d(u) \}$ is computable in $O(m)$ operations $(m \geq n)$. Then

$$T_{CD} = O\left(\frac{m R_{[0]}}{\mu^{1/2} \epsilon^{1/2}} \sum_{i=1}^{n} \| A e_i \| \right) \quad \leq \quad T_{FGM} = O\left(\frac{m n R_{[0]}}{\mu^{1/2} \epsilon^{1/2}} \| A \| \right).$$

It can be that $\| A e_i \| << \| A \|$, $i \in \{1 : n\}$.

Example: $0 < \gamma_1 \leq A^{(i,j)} \leq \gamma_2$, $\quad i \in \{1 : m\}$, $j \in \{1 : n\}$.

Then $\| A e_i \| \leq \gamma_2 \sqrt{m}$, and $\| A \| \geq \gamma_1 \| 1_m 1_n^T \| = \gamma_1 \sqrt{mn}$.

We gain $O(\sqrt{n})$ in the complexity.

Conclusion

1. Provided that $T(\nabla_i f(x)) = \frac{1}{n} T(f)$, we get methods, which are always more efficient than usual FGM.

2. Sometimes the gain reaches $O(\sqrt{n})$.

3. We get better results on a problem class, which contains many applications of Smoothing Technique.

4. Our method is oracle/iteration balanced ($O(n+m)$ operations for both).

But:

5. Constrained minimization is not covered.

6. Numerical testing is needed.