First-Order Algorithms Converge Faster than $O(1/k)$ on Convex Problems

LEE Ching-pei

Joint work with Stephen J. Wright
Main Results

• Several fundamental first-order methods for smooth or regularized optimization possess a convergence rate of $o(1/k)$ on convex problems.

• Better than the best known rate of $O(1/k)$.

<table>
<thead>
<tr>
<th>Smooth optimization</th>
<th>Hilbert space</th>
<th>Euclidean space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient descent</td>
<td>Coordinate descent</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Regularized optimization</th>
<th>Proximal gradient</th>
<th>Proximal coordinate descent</th>
</tr>
</thead>
</table>

• The key elements:
  - Descent method
  - Summability of $f(x_k) - f^*$ from an implicit regularization on the iterate distance to the solution set.

• The implicit regularization is algorithm-specific.
Gradient Descent

• Consider the following problem in a Hilbert space

\[ \min_x f(x), \]

with the solution set \( \Omega \) nonempty and \( f^* := \min_x f(x) \)

• \( f \) is \( L \)-Lipschitz continuously differentiable (called smooth from now on) and convex

• \( x_{k+1} \leftarrow x_k - \alpha_k \nabla f(x_k) \) with \( \alpha_k \) such that for given \( \gamma \in (0, 1] \)
  \( \alpha_{\text{max}} \geq \alpha_{\text{min}}, \) and \( \alpha_{\text{min}} \in (0, (2 - \gamma)/L], \)

\[
\begin{align*}
\alpha_k & \in [\alpha_{\text{min}}, \alpha_{\text{max}}], \\
f \left( x_k - \alpha_k \nabla f(x_k) \right) & \leq f(x_k) - \frac{\gamma \alpha_k}{2} \| \nabla f(x_k) \|^2
\end{align*}
\]

• Includes fixed step variants

• Best known existing convergence rate for \( f(x_k) - f^* \) is \( O(1/k) \), and we show a \( o(1/k) \) convergence rate
• Consider $\mathbb{R}^n$ ($n < \infty$) with the unit vectors $\{e_1, \ldots, e_n\}$, and the function $f$ has componentwise Lipschitz constants $L_1, \ldots, L_n > 0$ such that

$$|\nabla_i f(x) - \nabla_i f(x + he_i)| \leq L_i |h|, \quad \text{for all } x \in \mathbb{R}^n \text{ and all } h \in \mathbb{R}$$

• Given $\{\bar{L}_i\}_{i=1}^n$ such that $\bar{L}_i \geq L_i$ for all $i$, the CD update is

$$x_{k+1} \leftarrow x_k - \frac{\nabla_{i_k} f(x_k)}{\bar{L}_{i_k}} e_{i_k},$$

where $i_k$ is the coordinate selected for updating at the $k$th iteration.
• Stochastic coordinate descent (SCD) picks each $i_k$ independently following a pre-specified fixed probability distribution for all iterations:

$$p_i > 0, \quad i = 1, 2, \ldots, n; \quad \sum_{i=1}^{n} p_i = 1 \quad (1)$$

• Known similar $O(1/k)$ convergence rates to $f^*$ for $\mathbb{E}[f(x_k)]$ (expectation over the coordinate picks):
  1. Nesterov (2012) for $p_i \propto \bar{L}_i^\beta$ with $\beta \in [0, 1]$
  2. Qu and Richtárik (2016) for arbitrary sampling strategies satisfying (1)

• We get the same improvement to $o(1/k)$ for SCD with any samplings satisfying (1)
Consider regularized optimization of the form:

$$\min_x F(x) := f(x) + \Psi(x)$$

- $f$ smooth and convex as above,
- $\Psi$: convex, extended-valued, proper, and closed, can be nondifferentiable
Proximal Gradient

- **Proximal gradient** (Bruck Jr., 1975): \( x_{k+1} \leftarrow x_k + d_k \), where

\[
\begin{align*}
  d_k &= \arg\min_d \langle \nabla f(x_k), d \rangle + \frac{1}{2\alpha_k} \|d\|^2 + \psi(x_k + d), \\
  \alpha_k &\in [\alpha_{\min}, \alpha_{\max}], \quad F(x_k + d_k) \leq F(x_k) - \frac{\gamma}{2\alpha_k} \|d_k\|^2
\end{align*}
\]

- Known: in Hilbert spaces, the same \( O(1/k) \) convergence rate as gradient descent when \( f \) is convex

- We again get a \( o(1/k) \) convergence rate
Proximal Coordinate Descent

- Assume:
  - Euclidean space
  - $\Psi$ is separable: for $z = (z_1, \ldots, z_n)$, $\Psi(z) = \sum_{i=1}^{n} \Psi_i(z_i)$

- Extended from proximal gradient: like the extension from GD to CD:
  \[
  x_{k+1} \leftarrow x_k + d_{ik}^k e_i, \\
  d_{ik}^k := \arg\min_{d \in \mathbb{R}} \nabla_i f(x_k) d + \frac{\bar{L}_{ik}}{2} d^2 + \psi_i ((x_k)_{ik} + d)
  \]

- Known $O(1/k)$ convergence rates for convex $f$:
  - Lu and Xiao (2015): uniform sampling
  - Lee and Wright (2018): any sampling, with the additional assumption
  \[
  \max_{x: F(x) \leq F(x_0)} \text{dist}(x, \Omega) < \infty
  \]

- Again we extend the rate to $o(1/k)$ for any fixed sampling strategies, without any additional assumptions
See you at poster: Pacific Ballroom #207