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

LEE Ching-pei

WISCONSIN

WISCONSIN INSTITUTE FOR DISCOVERY

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 $O(1/k)$

	Hilbert space	Euclidean space
Smooth optimization	Gradient descent	Coordinate descent
Regularized optimization	Proximal gradient	Proximal coordinate descent

- 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

$$\min_{x} \quad f(x),$$

with the solution set Ω 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 α_k such that for given $\gamma \in (0, 1]$ $\alpha_{\max} \geq \alpha_{\min}$, and $\alpha_{\min} \in (0, (2-\gamma)/L]$,

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

- 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

Extension to Coordinate Descent

- Consider $\Re^n$ ($n < \infty$) 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

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

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

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

where i_k is the coordinate selected for updating at the kth iteration

Stochastic Coordinate Descent

- Stochastic coordinate descent (SCD) picks each i_k independently following a pre-specified fixed probability distribution for all iterations:

$$p_i > 0, \quad i = 1, 2, \dots, n; \quad \sum_{i=1}^{n} p_i = 1 \tag{1}$$

- 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)

Regularized Optimization

- Consider regularized optimization of the form:

$$\min_{x} \quad F(x) := f(x) + \Psi(x)$$

- f smooth and convex as above,
- Ψ: 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

$$\begin{cases} d_k = \operatorname{argmin}_d \, \langle \nabla f(x_k), d \rangle + \frac{1}{2\alpha_k}\|d\|^2 + \Psi(x_k + d), \\ \alpha_k \in [\alpha_{\min}, \alpha_{\max}], \quad F(x_k + d_k) \le F(x_k) - \frac{\gamma}{2\alpha_k}\|d_k\|^2 \end{cases}$$

- 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
  - Ψ 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:

$$x_{k+1} \leftarrow x_k + d_{i_k}^k e_{i_k},$$

$$d_{i_k}^k := \underset{d \in \Re}{\operatorname{argmin}} \; \nabla_{i_k} f(x_k) d + \frac{\bar{L}_{i_k}}{2} d^2 + \psi_{i_k}\left((x_k)_{i_k} + d\right)$$

- 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

$$\max_{x: F(x) \leq F(x_0)} \quad \mathrm{dist}(x, \Omega) < \infty$$

- Again we extend the rate to $o(1/k)$ for any fixed sampling strategies, without any additional assumptions

Thanks for Listening

See you at poster: Pacific Ballroom #207