Journées SMAI MODE 2018

Redémarrer la méthode de descente par coordonnée accélérée avec une estimation imprécise de la borne d'erreur quadratique

Olivier FERCOQ
Télécom-ParisTech – Université Paris Saclay

Zheng QU
The University of Hong Kong

Résumé. Nous proposons de nouvelles stratégies de redémarrage pour la méthode de descente par coordonnée accélérée. Notre contribution principale est de montrer que pour une suite bien choisie d'instants de redémarrage, la méthode redémarrée a un taux de convergence presque géométrique. Une caractéristique majeure de la méthode est qu'elle profite de la borne d'erreur quadratique locale sans utiliser explicitement sa valeur effective. Nous montrons aussi que sous l'hypothèse plus restrictive de forte convexité, un redémarrage à période fixe donne une taux de convergence géométrique, quelle que soit la période. Nous illustrons les performances de l'algorithme sur un problème de régression logistique et sur un problème de Lasso.

Mots-clefs : Descente par coordonnée accélérée; redémarrage; paramètre de forte convexité inconnu; borne d'erreur quadratique locale.

1 Motivation

We consider the minimization of composite convex functions of the form

$$F(x) = f(x) + \psi(x), \quad x \in \mathbb{R}^n$$

where f is differentiable with Lipschitz gradient and ψ may be nonsmooth but is separable, and has an easily computable proximal operator. Coordinate descent methods are often considered in this context thanks to the separability of the proximal operator of ψ. These are optimization algorithms that update only one coordinate of the vector of variables at each iteration, hence using partial derivatives rather than the whole gradient.
Similarly to what he had done for the gradient method, Nesterov introduced, for smooth functions, the randomized accelerated coordinate descent method with an improved guarantee on the iteration complexity [9]. Indeed, for a mild additional computational cost, accelerated methods transform the proximal coordinate descent method, for which the optimality gap $F(x_k) - F^\star$ decreases as $O(1/k)$, into an algorithm with "optimal" $O(1/k^2)$ complexity [8]. Lee and Sidford [3] introduced an efficient implementation of the method and Fercoq and Richtárik [2] developed the accelerated parallel and proximal coordinate descent method (APPROX) for the minimization of composite functions.
When solving a strongly convex problem, the classical (non-accelerated) gradient and coordinate descent methods automatically have a linear rate of convergence, i.e. $F(x_k) - F^\star \in O((1-\mu)^k)$ for a problem dependent $0 < \mu < 1$, whereas one needs to know explicitly the

strong convexity parameter in order to set accelerated gradient and accelerated coordinate descent methods to have a linear rate of convergence, see for instance [3, 4, 5, 9, 10]. Setting the algorithm with an incorrect parameter may result in a slower algorithm, sometimes even slower than if we had not tried to set an acceleration scheme [11]. This is a major drawback of the method because in general, the strong convexity parameter is difficult to estimate.
In the context of accelerated gradient method with unknown strong convexity parameter, Nesterov [10] proposed a restarting scheme which adaptively approximate the strong convexity parameter. The same idea was exploited in [6] for sparse optimization. Nesterov [10] also showed that, instead of deriving a new method designed to work better for strongly convex functions, one can restart the accelerated gradient method and get a linear convergence rate. It was later shown in [7, 1] that a local quadratic error bound is sufficient to get a global linear rate of convergence.

2 Contributions

We show how restarting the accelerated coordinate descent method can help us take profit of the local quadratic error bound of the objective, when this property holds.
We consider three setups:

1. If the local quadratic error bound coefficient μ of the objective function is known, then we show that setting a restarting period as $O(1/\sqrt{\mu})$ yields an algorithm with optimal rate of convergence. More precisely restarted APPROX admits the same theoretical complexity bound as the accelerated coordinate descent methods for strongly convex functions developed in [5], is applicable with milder assumptions and exhibits better performance in numerical experiments.

2. If the objective function is strongly convex, we show that we can restart the accelerated coordinate descent method at *any* frequency and get a linearly convergent algorithm. The rate depends on an estimate of the local quadratic error bound and we show that for a wide range of this parameter, one obtains a faster rate than without acceleration. In particular, we do not require the estimate of the error bound coefficient to be smaller than the actual value.

3. If the local error bound coefficient is not known, we introduce a variable restarting periods and show that up to a $\log(\log 1/\epsilon)$ term, the algorithm is as efficient as if we had known the local error bound coefficient.

Références

[1] Olivier Fercoq and Zheng Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *arXiv preprint arXiv:1709.02300*, 2017.

[2] Olivier Fercoq and Peter Richtárik. Accelerated, parallel and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

[3] Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 147–156. IEEE, 2013.

Algorithm 1 APPROX(f, ψ, x_0, K)

Set $\theta_0 = \frac{\tau}{n}$ and $z_0 = x_0$.
for $k \in \{0, \dots, K-1\}$ **do**
  $y_k = (1-\theta_k)x_k + \theta_k z_k$
  Randomly generate $S_k \sim \hat{S}$
  for $i \in S_k$ **do**
    $z_{k+1}^i = \arg\min_{z \in \mathbb{R}} \left\{ \langle \nabla_i f(y_k), z - y_k^i \rangle + \frac{\theta_k n v_i}{2\tau} |z - z_k^i|^2 + \psi^i(z) \right\}$
  end for
  $x_{k+1} = y_k + \frac{n}{\tau}\theta_k (z_{k+1} - z_k)$
  $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$
end for
return x_K

Algorithm 2 APPROX with restart

Choose $x_0 \in \operatorname{dom} \psi$ and set $\bar{x}_0 = x_0$.
Choose RestartTimes $\subseteq \mathbb{N}$.
for $r \geq 0$ **do**
  K = RestartTimes$(r+1)$ - RestartTimes(r)
  $\bar{x}_{r+1}$ = APPROX$(f, \psi, \bar{x}_r, K)$
end for

[4] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3384–3392. Curran Associates, Inc., 2015.

[5] Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated proximal coordinate gradient method. In *Advances in Neural Information Processing Systems*, pages 3059–3067, 2014.

[6] Qihang Lin and Lin Xiao. An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. *Computational Optimization and Applications*, 60(3):633–674, 2015.

[7] Mingrui Liu and Tianbao Yang. Adaptive accelerated gradient converging methods under holderian error bound condition. *Preprint arXiv:1611.07609*, 2016.

[8] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[9] Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[10] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[11] Brendan O'Donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, pages 1–18, 2012.