Stochastic Coordinate Descent for Nonsmooth Convex Optimization

Qi Deng
Department of CISE, University of Florida
qdeng@cise.ufl.edu

Jeffrey Ho
Department of CISE, University of Florida
jho.jeffrey@gmail.com

Anand Rangarajan*
Department of CISE, University of Florida
anand@cise.ufl.edu

Abstract

Stochastic coordinate descent, due to its practicality and efficiency, is increasingly popular in machine learning and signal processing communities as it has proven successful in several large-scale optimization problems , such as l1 regularized regression, Support Vector Machine, to name a few. In this paper, we consider a composite problem where the nonsmoothness has a general structure that is compatible with a coordinate partition, and we solve the nonsmooth optimization problem using a sequence of smooth approximations. In particular, we extend Nesterov's estimate sequence technique by incorporating smooth approximation and coordinate randomization. By studying the effect of smooth approximation, we develop rules for selecting smooth approximations that not only guarantee the algorithm's convergence but also provide better convergence rate than the subgradient black-box model. Specifically, we obtain the convergence rate of $O\left(\frac{1}{K}\right)$ for nonsmooth convex functions and $O\left(\frac{1}{K^2}\right)$ for strongly convex functions. The convergence analysis developed in this paper and the results, to the best of our knowledge, have not been shown previously for stochastic coordinate descent.

1 Introduction

In this paper, we are interested in applying stochastic coordinate descent to solve the optimization of a (convex) composite function $f(x) = h(x) + g(x)$ where the objective $f(x)$ contains two parts: $h(x)$ is differentiable with Lipschitz continuous gradient and $g(x)$ is a general convex nonsmooth function (satisfying conditions specified later). Although this problem has been investigated quite extensively in the literature, none of the earlier work has studied the problem from the perspective of stochastic coordinate descent. For huge-scale optimization problems that are increasingly common in machine learning and other application domains, coordinate descent is often the only available method due to its practicality, and accordingly, it is experiencing a resurgence of interest recently, e.g., [Nes10]. In this context, we propose an accelerated coordinate-descent scheme for minimizing the convex composite function $f(x)$ based on the idea of sequential coordinate smooth approximation. Specifically, we introduce a sequence $g_{\mu_k}(x)$ of increasingly-accurate smooth approximations of $g(x)$ such that the smoothing parameter μ_k indexed by the iteration counter k is a nonnegative real number converging to zero as $k \to \infty$. At each iteration, our scheme uses the smooth approximation $f_{\mu_k}(x) = h(x) + g_{\mu_k}(x)$ as the surrogate and with suitable assumptions on the form of g_{μ_k}, accelerated coordinate descent scheme can be developed for the smooth approximation sequence $f_k(x)$. Furthermore, the sequential smoothing can be incorporated into the estimate sequence framework of Nesterov for analyzing the convergence complexity, with the smoothing parameter μ_k playing the crucial role of balancing the degree of smoothing and the rate of convergence. We study different smoothing strategies in terms of using different sequences of μ_k, and with appropriately-chosen

*This work was partially supported by NSF IIS 1065081

smoothing parameters, the proposed scheme provides the (worst-case) complexity of $O\left(\frac{1}{K}\right)$ when f is convex, and $O\left(\frac{1}{K^2}\right)$ when f is strongly convex. To the best of our knowledge, this is the first theoretical convergence analysis of coordinate descent using the exact coordinate gradient and randomized coordinate selection on nonsmooth problems.

Related Work: When the non-smooth component $g(x)$ is absent or possesses a simple structure (such as in the lasso objective), Nestervo's celebrated accelerated gradient methods [Nes83], [Nes03], [Nes12], [BT09] are known to be optimal in terms of the worst case complexity. In general, the addition of a nonsmooth convex function makes the convergence complexity worse than in the corresponding smooth case, and only subgradient methods have been known to converge. A well-known technique for handling a nonsmooth function is by solving a modified problem using its smooth approximation. In [Nes05b], it is shown that if $g(x)$ has a certain structure, optimization of $f(x)$ can still take advantage of fast gradient schemes on the approximated smooth problem, with an achievable convergence rate that is better than the subgradient method by an order of magnitude. Smooth approximation in the primal-dual context is also studied in [Nes05a], providing further speed-ups for strongly convex functions. The work in [BT12] further extends the realm of "smoothable" functions, and it demonstrates the benefit of smooth approximations (with a fast gradient scheme) for several interesting problems. However, we remark that these first-order methods require the full-knowledge of the gradient and therefore, they and their convergence guarantees do not admit an immediate extension to methods using coordinate descent. Furthermore, smooth approximation does not seem to have been considered from the sequential viewpoint (as we do here), except perhaps in [Nes05a] where a fixed sequence of smoothing parameters μ_k is suggested.

2 Algorithm and Convergence Analysis

The optimization problem of interest is

$$\min_{x\in\mathcal{X}} f(x) = h(x) + g(x). \tag{1}$$

Our focus will be on unconstrained problems with $\mathcal{X} = \mathbb{R}^d$ for some $d > 0$. Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_p$ denote a Cartesian decomposition of $\mathcal{X}$, and it correspondingly defines a partition of the coordinates of $\mathcal{X}$. For notational simplicity, we will use $\mathcal{X}_i$ to denote both the Cartesian factor and its associated coordinate block. In (1), $h(x)$ is differentiable and σ-strongly convex with respect to the Euclidean norm. [We should emphasize that in general we don't assume strong convexity of h, so σ (the constant in front of $\|x-x'\|^2$) can be 0.]. Furthermore, we assume that the gradient of $h(x)$ is Lipschitz continuous with constant L_h and its partial gradient $\nabla_i h$ with respect to the coordinate block $\mathcal{X}_i$ is also Lipschitz continuous with constant L_{h_i}. To facilitate the analysis, we assume the nonsmooth component $f(x)$ can be approximated by smooth functions using the following notion:

Definition 1 (Sequential Block Coordinate Smooth Approximation:) *Let $\mu = \{\mu_1, ...\mu_k, ..\}$ denote a decreasing sequence of non-negative real numbers such that $\mu_k \to 0$ as $k \to \infty$ and $\mathbf{B} = \{\beta_1, ..., \beta_p\}, \mathbf{L} = \{L_1, L_2, \ldots, L_p\}$, two sets of p positive real numbers. We say the function f is $(\delta, \mathbf{B}, \mathbf{L})$-block coordinate smoothable, and $\{f_{\mu_k}\}$ is a sequential Block Coordinate Smoothing Approximation of $f(x)$ if the following conditions are satisfied for some $\delta > 0$:*

1. *(Monotonicity) For each $x \in \mathcal{X}$, $f_{\mu_1}(x), f_{\mu_2}(x), ...$ form a decreasing sequence with a lower bound given by $f(x)$. Specifically, $\forall x \in \mathcal{X}$, $f_{\mu_1}(x) \geqslant f_{\mu_2}(x) \geqslant \ldots \geqslant f_{\mu_k}(x) \geqslant \ldots \geqslant f(x)$.*
2. *$f_\mu(x) \leqslant f(x) + \delta\mu$, for every $x \in X$, μ.*
3. *For each f_{μ_k}, its partial gradient $\nabla_i f_{\mu_k}$ with respect to the coordinate block $\mathcal{X}_i$ is Lipschitz continuous with constant $L_i^{(k)} = L_i + \frac{\beta_i}{\mu_k}$.*

We note that since the sequence $\{\mu_k\}$ is decreasing, f_{μ_k} converges uniformly to f on any compact subset of $\mathcal{X}$ [also by (2)], and the sequence f_{μ_k} always approximates f from above. This peculiar feature is required for the later convergence analysis using Nesterov's estimate sequence technique, and in this regard, our smoothing scheme is different from those discussed in [BT12] that only require f_{μ_k} to lie within a band around f. In particular, it can also be considered as a special extension of the smoothing paradigm discussed in [BT12] to block-wise coordinate setting.

An Example: A quick example of $(\delta, \mathbf{B}, \mathbf{L})$ -sequential block coordinate smooth approximation of $f(x)$ can be constructed from the smoothing example given in [Nes05b]. In this example, the

nonsmooth function $g(x)$ assumes a special form $g(x) = \max_{u\in\mathcal{U}}\{\langle Ax, u\rangle - \phi(u)\}$ for some linear map $A : \mathbb{R}^d \to \mathbb{R}^q$ and a convex function $\phi(u)$ defined on a bounded subset $\mathcal{U} \subset \mathbb{R}^q$. This particular type of g appears in a variety of problems that are important in machine learning and signal processing, e.g, the well-known hinge loss (SVM), quantile regression and TV-norm denoising. By Dansker's theorem, nonsmoothness of $g(x)$ is essentially equivalent to the non-uniqueness of the global optimal solution of $\max_{u\in\mathcal{U}}\{\langle Ax, u\rangle - \phi(u)\}$. Therefore, a simple smoothing method is to introduce an additional strongly convex term to ensure the uniqueness of the optimal solution: $\tilde{g}_\mu = \max_{u\in\mathcal{U}}\{\langle Ax, u\rangle - \phi(u) - \mu d(u)\}$, with the strongly convex function $d(u)$ satisfying $d(u) \geqslant \frac{1}{2}\|u\|^2$. Define $\tilde{f}_\mu(x) = h(x) + \tilde{g}_\mu(x)$. It follows that the gradient of $\tilde{f}_\mu$ is Lipschitz continuous with constant $L_h + \frac{\|A\|^2}{\mu}$. To account for the monotonicity condition, we modify $\tilde{f}_\mu$ by adding a positive constant linear in μ:

$$f_\mu = h(x) + \max_{u\in\mathcal{U}}\{\langle Ax, u\rangle - \phi(u) - \mu\left[d(u) - \max_{v\in\mathcal{U}} d(v)\right]\}. \tag{2}$$

The extra term does not change the differentiability of f_μ, but ensures the monotonicity of $f_\mu(\cdot)$ as a function of μ for $\mu > 0$, In addition, it can be shown that $f_\mu(x) \leq f(x) + \mu \max_{v\in\mathcal{U}} d(v)$. More importantly, because of the above maximization structure, we can exploit the differentiability of f_μ coordinate-wise. In particular, for a decreasing sequence $\mu = \{\mu_1, \mu_2, ...\}$ and $\beta_i = \|A\|_{1i,2} = \max_{x,y}\{\langle A_i x, y\rangle, \|x\|_{\mathcal{X}_i} = 1, \|y\|_{\mathcal{U}} = 1, x \in \mathcal{X}, y \in \mathcal{U}\}, i = 1, ..., p$, the functions f_{μ_k} form a $(\delta, \mathbf{B}, \mathbf{L})$-sequential block coordinate smooth approximation of f.

Optimization Algorithm: The algorithm presented here is a simple and efficient extension of Nesterov's method in [Nes10] that applies the accelerated first-order method to coordinate descent, and the detail of the algorithm is outlined below. We use f_k as a shorthand notation for f_{μ_k}, and L_i for L_{h_i},$L_i^{(k)} = L_i + \frac{\beta_i}{\mu_k}$. In coordinate selection, we adopt stochastic sampling of coordinates with probability proportional to the scale of the corresponding (estimated) Lipschitz constant of f_k: $\{p_i = L_i^{(k)}/S^{(k)}\}$, where $S^{(k)}$ is the normalization factor. The extension to other sampling schemes such as uniform sampling will be considered in future work. The algorithm maintains two sequences of variables $\{x_k\}, \{y_k\}$, $\{x_k\}$ is updated using the block coordinate gradient on some randomly sampled coordinate of y_k. Meanwhile, $\{y_k\}$ is interpolated from the previous updates, controlled by a set of carefully chosen parameters $\{\alpha_k, \gamma_k\}$. This is in the spirit of Nesterov's accelerated gradient method. In addition, since f is nonsmooth, the block coordinate gradient of the smooth surrogate function f_k is applied to construct the convergence sequence. As a trade-off, the error of the smooth approximation affects the convergence rate and it. explains why the convergence rate is worse than results for optimizing smooth functions. Recent work in [Nes10] [LX13], [LS13] extend the concept of estimate sequence to randomized settings, to accelerate coordinate descent. In the context of these very recent developments, our work can be considered as an extension that incorporates smoothing techniques under the coordinate descent framework. The main difference is that we maintain a non-increasing sequence of positive real variables $\{\mu_k\}_{k=1,2,...K,}$ and use the associated sequence of smooth functions $\{f_{\mu_k}\}_k$ to approximate f in a gradually-more-accurate manner. The convergence analysis of the algorithm in terms of the sequence $\{\mu_k\}$ and the expected value $Ef(x_{K+1})$ at $K+1$-step is summarized in the following theorem (assuming an appropriately chosen $(\delta, \mathbf{B}, \mathbf{L})$-block coordinate smoothing sequence).

Theorem 1 *For the optimization problem defined in (1), if $f(x)$ is a non-strongly convex function with $\sigma = 0$, let $\alpha_k = \frac{2}{k+3}$, $\mu_k = \frac{c}{(k+2)}$ for $c = \sqrt{\frac{2p\beta_0}{\delta}}\|x_0 - x^*\|$, $\gamma_0 = 4p\left(L_0 + \frac{\beta_0}{c}(K+2)\right)$, we have*

$$E\left[f(x_{K+1})\right] - f^* \leq \frac{2\sqrt{2p\delta\beta_0}\|x_0 - x^*\|}{(K+3)} + \frac{2\left(\phi_0(x_0) - f(x^*)\right) + 4pL_0\|x_0 - x^*\|^2}{(K+3)(K+2)}.$$

If $h(x)$ is a strongly convex function with $\sigma > 0$, and let $\alpha_k = \frac{3}{k+4}$, $c = \frac{18p\beta_0}{\sigma}$, $\mu_k = \frac{c}{(k+3)(k+2)}$, and $\gamma_0 = 3pL_0(K+2) + \sigma$, we have

$$E\left[f(x_{K+1})\right] - f^* \leq \frac{54\delta p\beta_0}{\sigma(K+4)(K+3)} + \frac{9pL_0\|x_0 - x^*\|^2}{(K+4)(K+3)} + \frac{6\left[\phi_0(x_0) - f(x^*)\right] + 3\sigma\|x_0 - x^*\|^2}{(K+4)(K+3)(K+2)}.$$

The convergence analysis presented in the theorem is based on Nesterov's method of estimate sequences [Nes03][LX13]. A crucial estimate in the analysis is the inequality $Ef(x_{K+1}) - f(x^*) \leq$

Algorithm 1: Stochastic Coordinate Smooth Approximation

Input : A $(\delta, \mathbf{B}, \mathbf{L})$-block-coordinate smoothable function f and K the total number of iterations. Initialize γ_0, α_0, $L_0 = \sum_{j=1}^{p} L_j$ and $\beta_0 = \sum_i^p \beta_i$, x_0, y_0.

for $k = 0, 1, 2, \ldots, K$ **do**

Set μ_k (as a non-decreasing sequence) and its associated smooth approximation $f_k(\cdot)$.

For $j = 1, \ldots, p$, set $L_j^{(k)} = L_j + \frac{\beta_j}{\mu_k}$, $S^{(k)} = \sum L_j^{(k)}$, $L^{(k)} = [S^{(k)}]^2 / \min_j L_j^{(k)}$.

Sample $i \in \{1, 2, \ldots p\}$ with prob$(i = j) = L_j^{(k)} / S^{(k)}$.

Choose γ_{k+1}, α_k such that $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k \sigma \geq L^{(k)} \alpha_k^2$ and let $\theta_{k+1} = \frac{\alpha_{k+1}\gamma_{k+1}}{\gamma_{k+1} + \alpha_{k+1}\sigma}$.

Set

$$x_{k+1} = y_k - \frac{1}{L_i} \nabla_i f_k(y_k), \tag{3}$$

$$y_{k+1} = x_{k+1} + \theta_{k+1} \left(\frac{1 - \alpha_k}{\alpha_k} (x_{k+1} - x_k) + \frac{1}{\alpha_k L_i^{(k)}} \left(1 - \frac{\alpha_k^2 S^{(k)}}{\gamma_{k+1}} \right) \nabla_i f_k(y_k) \right). \tag{4}$$

output: x_{K+1}

$\lambda_{K+1} (\phi_0(x_0) - f(x^*))$ that bounds the gap $Ef(x_{K+1}) - f(x^*)$ from above (where $\phi_0(x)$ is a function in $\mathcal{X}$ that can be set to f_0), and the convergence rate can be extracted from the sequence λ_k. For our analysis, the upper bound on the gap now acquires an important residual term Δ_{K+1}, $Ef(x_{K+1}) - f(x^*) \leq \lambda_{K+1} (\phi_0(x^*) - f(x^*)) + \Delta_{K+1}$, that depends on the smoothing parameters $\{\mu_k\}$. The analysis is centered on balancing the trade off betweenΔ_{K+1} and λ_{K+1}in showing that — with an appropriately-chosen sequence $\{\mu_k\}$ — the desired convergence rate can still be extracted from the sequence λ_k.

For non-strongly convex functions, our complexity bound is dominated by the $O(\frac{1}{K})$-order term that is essentially equivalent to the asymptotic error (approximation accuracy) of the smoothing sequence with Lipschitz constant β_0. For strongly convex functions, the additional information σ provides a more accurate approximation with error in the order of $O(\frac{1}{K^2})$, and by our choice of parameters, the optimization on the smooth part is also of the same order. We remark that our complexity result agrees with the result in [Nes05a] but is proved in a more general context without explicit dependence on duality. A related but different approach is presented in [BT12] where the authors consider the class of smoothable functions and claim that for convex nonsmooth problems with a fixed smooth approximation (using a smoothing constant μ depending on the total iteration number K), one can always apply any accelerated scheme as the black-box toolbox to obtain accelerated convergence over the subgradient method. However, our analysis together with Nesterov's result [Nes05a] suggest that this approach may be suboptimal (yielding only $\Omega(\frac{1}{K^2})$) for strongly convex functions with an adaptive smoothing sequence needed to attain the faster convergence of $O(\frac{1}{K^2})$. More precisely, let $\mu \equiv \frac{1}{K^\gamma}$ for some $\gamma < 2$. The approximation error of the smoothing is asymptotically $O\left(\frac{1}{K^\gamma}\right)$. On the other hand, if we fix $\mu_K = O(\frac{1}{K^2})$, applying a first-order accelerated scheme to the problem of minimizing f_μ cannot go beyond the complexity rate of $O\left(\left(1 - \sqrt{\frac{\sigma}{L}}\right)^K\right)$. In this case, the associated Lipschitz constant is $L = O(\frac{1}{\mu}) = O(K^2)$ and the optimization error for f_μ is asymptotically on the order of

$$\lim_{K \to \infty} \left(1 - O\left(\frac{1}{K}\right)\right)^K = O(1). \tag{5}$$

This suggests that when a smooth function with a large Lipschitz constant (i.e., a nearly-singular smooth function) is used as an approximation, even accelerated methods may not guarantee convergence in the first K iterations.

3 Conclusion

In this paper, we have employed smoothing techniques within the coordinate descent framework and coupled it to Nesterov's accelerated scheme to significantly improve on the black-box subgradient model. This results in an efficient stochastic coordinate descent algorithm for optimizing nonsmooth convex functions. Our approach is immediately applicable to well-known optimization problems involving the SVM hinge loss, quantile regression, TV-norm denoising and others. In the immediate future, we plan to specialize our approach to obtain efficient algorithms for solving these and other similar nonsmooth convex optimization problems in machine learning and image processing.

References

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[BT12] A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

[LS13] Y. T. Lee and A. Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. *arXiv preprints arXiv:1305.1922*, 2013.

[LX13] Z. Lu and L. Xiao. On the Complexity Analysis of Randomized Block-Coordinate Descent Methods. *arXiv preprints arXiv:1305.4723*, 2013.

[Nes83] Y. Nesterov. A method of solving a convex programming problem with convergence rate $\mathrm{O}(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[Nes03] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2003.

[Nes05a] Y. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM J. on Optimization*, 16(1):235–249, 2005.

[Nes05b] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1), May 2005.

[Nes10] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. CORE Discussion Papers 2010002, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2010.

[Nes12] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, pages 1–37, 2012.