Stochastic Coordinate Minimization with Progressive Precision for Stochastic Convex Optimization

Sudeep Salgia [1] **Qing Zhao** [1] **Sattar Vakili** [2]

Abstract

A framework based on iterative coordinate minimization (CM) is developed for stochastic convex optimization. Given that exact coordinate minimization is impossible due to the unknown stochastic nature of the objective function, the crux of the proposed optimization algorithm is an optimal control of the minimization precision in each iteration. We establish the optimal precision control and the resulting order-optimal regret performance for strongly convex and separably nonsmooth functions. An interesting finding is that the optimal progression of precision across iterations is independent of the low-dimensional CM routine employed, suggesting a general framework for extending low-dimensional optimization routines to high-dimensional problems. The proposed algorithm is amenable to online implementation and inherits the scalability and parallelizability properties of CM for large-scale optimization. Requiring only a sublinear order of message exchanges, it also lends itself well to distributed computing as compared with the alternative approach of coordinate gradient descent.

1. Introduction

1.1. Stochastic Convex Optimization

Stochastic convex optimization aims at minimizing a random loss function $F(\mathbf{x};\xi)$ in expectation:

$$f(\mathbf{x}) = \mathbb{E}_{\xi}\left[F(\mathbf{x};\xi)\right], \tag{1}$$

where $\mathbf{x}$ is the decision variable in a convex and compact set $\mathcal{X} \subset \mathbb{R}^d$ and ξ is an endogenous random vector. The probabilistic model of ξ is unknown, or even when it is known, the expectation of $F(\mathbf{x};\xi)$ over ξ cannot be analytically characterized. As a result, the objective function $f(\mathbf{x})$ is unknown.

With the objective function unknown, the decision maker can only take a trial-and-error learning approach by choosing, sequentially in time, a sequence of query points $\{\mathbf{x}_t\}_{t=1}^T$ with the hope that the decisions improve over time. Various error feedback models have been considered. The zeroth-order vs. first-order feedback pertains to whether the random loss $F(\mathbf{x}_t;\xi_t)$ or its gradient $G(\mathbf{x}_t;\xi_t)$ at each query point $\mathbf{x}_t$ is used in the learning algorithm. The full-information vs. bandit feedback relates to whether the entire loss function $F(\mathbf{x};\xi_t)$ over all $\mathbf{x}$ or only the random loss/ gradient at the queried point $\mathbf{x}_t$ is revealed at each time.

The performance measure has traditionally focused on the convergence of $\mathbf{x}_T$ to the minimizer $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ or $f(\mathbf{x}_T)$ to $f(\mathbf{x}^*)$. In an online setting, a more suitable performance measure is the cumulative regret defined as the expected cumulative loss at the query points in excess to the minimum loss:

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T}(F(\mathbf{x}_t;\xi_t) - f(\mathbf{x}^*))\right]. \tag{2}$$

This performance measure gives rise to the exploration-exploitation tradeoff: the need to explore the entire domain $\mathcal{X}$ for the sake of future decisions and the desire to exploit the currently best decision indicated by past observations to reduce present loss. A learning algorithm with a sublinear regret order in T implies the convergence of $f(\mathbf{x}_T)$ to $f(\mathbf{x}^*)$, and the specific order measures the rate of convergence.

The archetypal statistical learning problem of classification based on random instances is a stochastic optimization problem, where the decision variable $\mathbf{x}$ is the classifier and ξ the random instance consisting of its feature vector and hidden label. The probabilistic dependence between feature and label is unknown. Another example is the design of large-scale complex physical systems that defy analytical modeling. Noisy observations via stochastic simulation is all that is available for decision making.

[1]School of Electrical & Computer Engineering, Cornell University, Ithaca, NY [2]Prowler.io, Cambridge, UK. Correspondence to: Sudeep Salgia <ss3827@cornell.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

1.2. From SGD to SCD

Stochastic convex optimization was pioneered by Robbins and Monro in 1951 (Robbins & Monro, 1951), who studied the problem of approximating the root of a monotone function $g(\mathbf{x})$ based on successive observations of noisy function values at chosen query points. The problem was originally referred to as stochastic approximation, later also known as stochastic root finding (Pasupathy et al., 2011). Its equivalence to the first-order stochastic convex optimization is immediate when $g(\mathbf{x})$ is viewed as the gradient of a convex function $f(\mathbf{x})$ to be minimized. The stochastic gradient descent (SGD) approach developed by Robbins and Monro (Robbins & Monro, 1951) has long become a classic and is widely used. The basic idea of SGD is to choose the next query point $\mathbf{x}_{t+1}$ in the opposite direction of the observed gradient while ensuring $\mathbf{x}_{t+1} \in \mathcal{X}$ via a projection operation. Numerous variants of SGD with improved performance have since been developed and their performance analyzed under various measures (See (Ruder, 2016; Bottou et al., 2018) for recent surveys).

The high cost in computing full gradients in large-scale high-dimensional problems and the resistance of SGD to parallel and distributed implementation have prompted the search for alternative approaches that enjoy better scalability and parallelizability.

A natural choice is iterative coordinate minimization (CM) that has been widely used and analyzed for optimizing a known deterministic function (Wright, 2015). Also known as alternating minimization, CM is rooted in the methodology of decomposing high-dimensional problems into a sequence of simpler low-dimensional ones. Specifically, CM-based algorithms approach the global minimizer by moving successively to the minimizer in each coordinate[1] while keeping other coordinates fixed to their most recent values. For known deterministic objective functions, it is often assumed that the minimizer in each coordinate can be computed, and hence attained in each iteration.

When coordinate-wise minimization is difficult to carry out, coordinate (gradient) descent (CD) can be employed, which takes a single step (or a fixed number of steps) of (gradient) descent along one coordinate and then moves to the next coordinate[2]. For quadratic objective functions, CD with properly chosen step sizes essentially carries out coordinate minimization. For general objective functions, however, it is commonly observed that CM outperforms CD (Wright, 2015; Beck & Tetruashvili, 2013; Tibshirani, 2013).

While CD/CM-based algorithms have been extensively studied for optimizing deterministic functions, their extensions and resulting performance for stochastic optimization are much less explored. CD can be applied to stochastic optimization with little modification. Since the noisy partial gradient along a coordinate can be viewed as an estimate of the full gradient, stochastic coordinate descent (SCD) has little conceptual difference from SGD. In particular, when the coordinate is chosen uniformly at random at each time, the noisy partial gradient along the randomly chosen coordinate is an unbiased estimate of the full gradient. All analyses of the performance of SGD directly apply. More sophisticated extensions of CD-based methods have been developed in a couple of recent studies (see Sec. 1.4).

Since exact minimization along a coordinate is impossible due to the unknown and stochastic nature of the objective function, the extension of CM to stochastic optimization is much less clear. This appears to be a direction that has not been taken in the literature and is the focus of this work.

1.3. Main Results

While both CD- and CM-based methods enjoy scalability and parallelizability, CM often offers better empirical performance and has a much lower overhead in message exchange in distributed computing (due to its sublinear order of switching across coordinates in comparison to the linear order in CD). It is thus desirable to extend these advantages of CM to stochastic optimization.

In this paper, we study stochastic coordinate minimization (SCM) for stochastic convex optimization. We develop a general framework for extending any given low-dimensional optimization algorithm to high-dimensional problems while preserving its level of consistency and regret order. Given that exact minimization along coordinates is impossible, the crux of the proposed framework—referred to as Progressive Coordinate Minimization (PCM)—is an optimal control of the minimization precision in each iteration. Specifically, a PCM algorithm is given by a tuple $(\{\epsilon_k\}, \upsilon, \tau)$, where $\{\epsilon_k\}_{k\in\mathbb{N}}$ governs the progressive precision of each CM iteration indexed by k, υ is an arbitrary low-dimensional optimization routine employed for coordinate minimization, and τ is the self-termination rule for stopping υ at the given precision ϵ_k in each iteration k. We establish the optimal precision control and the resulting order-optimal regret performance for strongly convex and separably non-smooth functions. An interesting finding is that the optimal progression of precision across iterations is independent of the low-dimensional routine υ, suggesting the generality of the framework for extending low-dimensional optimization

[1] We use the term "coordinate" to also refer to a block of coordinates.

[2] The term coordinate descent is often used to include coordinate minimization. We make an explicit distinction between CD and CM in this paper. The former refers to taking a single step (or a pre-fixed number of steps) of (gradient) descent along one coordinate and then move to another coordinate. The latter moves along each coordinate with the specific goal of arriving at the minimizer (or a small neighborhood) in this coordinate before switching to another coordinate.

algorithms to high-dimensional problems.

We also illustrate the construction of order-optimal termination rules for two specific optimization routines: SGD (applied to minimize along a coordinate) and RWT (recently proposed in (Vakili & Zhao, 2019; Vakili et al., 2019)). While SGD is directly applicable to high-dimensional problems, its extension within the PCM framework leads to a marriage between the efficiency of SGD with the scalability and parallelizability of CM. RWT as proposed in (Vakili & Zhao, 2019; Vakili et al., 2019) is only applicable to one-dimensional problems. With no hyper-parameters to tune, however, it has an edge over SGD in terms of robustness and self-adaptivity to unknown function characteristics. For both low-dimensional routines, we demonstrate their high-dimensional extensions within the PCM framework. Empirical experiments using the MNIST dataset show superior performance of PCM over SCD, which echoes the comparison between CM and CD in deterministic settings.

1.4. Related Work

CD/CM-based methods for optimizing a known deterministic function have a long history. While such methods had often been eclipsed by more high-performing algorithms, they have started to enjoy increasing popularity in recent years due to the shifted needs from high accuracy to low cost, scalability, and parallelizability in modern machine learning and data analytics applications (Hsieh et al., 2008a;b; Nesterov, 2012; 2014; Richtárik & Takáč, 2016a;b). See (Wright, 2015; Fercoq & Richtárik, 2019) for a detailed literature survey with insights on the development of CD/CM methods over the years.

Early studies on the convergence of CM-based approaches include (Luo & Tseng, 1992; Tseng, 2001; Tseng & Yun, 2008; 2009; Saha & Tewari, 2013). CD-based methods have proven to be easier to analyze, especially under the setup of randomized selection of coordinates (Nesterov, 2012; Leventhal & Lewis, 2010; Tewari & Shalev-Shwartz, 2011; Tao et al., 2012; Deng et al., 2013; Shalev-Shwartz & Zhang, 2014; Csiba et al., 2015; Karimi et al., 2016; Salehi et al., 2018). Such CD/CM-based algorithms are often referred to as stochastic CD/CM in the literature due to the randomly chosen coordinates. Optimizing a known deterministic function, however, they are fundamentally different from the stochastic optimization algorithms considered in this work. The term CD/CM with random coordinate selection as used in (Richtárik & Takáč, 2014; Lu & Xiao, 2015) gives a more accurate description.

CD-based methods have been extended to mini batch settings or for general stochastic optimization problems (Razaviyayn et al., 2013; Wang & Banerjee, 2014; Zhao et al., 2014; Dang & Lan, 2015; Reddi et al., 2015; Xu & Yin, 2015; Zhang & Gu, 2016; Konečný et al., 2016). In particular, (Dang & Lan, 2015) extended block mirror descent to stochastic optimization. Relying on an averaging of decision points over the entire horizon to combat stochasticity, this algorithm is not applicable to online settings and does not seem to render tractable regret analysis. (Wang & Banerjee, 2014) gave an online implementation of SCD, which we compare with in Sec. 7.

The progressive precision control framework developed in this work bears similarity with inexact coordinate minimization that has been studied in the deterministic setting (see, for example, (Deng et al., 2013; Razaviyayn et al., 2013; Grippo & Sciandrone, 1999; Tappenden et al., 2016)). The motivation for inexact minimization in these studies is to reduce the complexity of the one-dimensional optimization problem, which is fundamentally different from the root cause arising from the unknown and stochastic nature of the objective function. The techniques involved hence are inherently different with different design criteria.

2. Problem Formulation

We consider first-order stochastic convex optimization with bandit feedback. The objective function $f(\mathbf{x})$ over a convex and compact set $\mathcal{X} \subset \mathbb{R}^d$ is unknown and stochastic as given in (1). Let $g(\mathbf{x}) \equiv \nabla f(\mathbf{x})$ be the (sub)gradient of $f(\mathbf{x})$. Let $G(\mathbf{x};\xi)$ denote unbiased gradient estimates satisfying $\mathbb{E}_\xi[G(\mathbf{x};\xi)] = g(\mathbf{x})$. Let $g_i(\mathbf{x})$ (similarly, $G_i(\mathbf{x};\xi)$) denote the partial (random) gradient along the i-th coordinate ($i = 1,\ldots,d$). Let $\mathbf{x}_i$ and $\mathbf{x}_{-i}$ denote, respectively, the i-th element and the $(d-1)$ elements other than the i-th element of $\mathbf{x}$. We point out that while we focus on coordinate-wise decomposition of the function domain, extension to a general block structure is straightforward.

2.1. The Objective Function

We consider objective functions that are convex and possibly non-smooth with the following composite form:

$$f(\mathbf{x}) = \psi(\mathbf{x}) + \phi(\mathbf{x}), \tag{3}$$

where $\phi(\mathbf{x})$ is a coordinate-wise separable convex function (possibly non-smooth) of the form $\phi(\mathbf{x}) = \sum_{i=1}^d \phi_i(\mathbf{x}_i)$ for some one-dimensional functions $\{\phi_i(x),\, x \in \mathcal{X}_i\}_{i=1}^d$ and ψ is α-strongly convex and β-smooth. More specifically, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$\psi(\mathbf{y}) \geq \psi(\mathbf{x}) + \langle \nabla\psi(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{\alpha}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \tag{4}$$

$$\|\nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{y})\| \leq \beta\|\mathbf{x}-\mathbf{y}\|_2 \tag{5}$$

Let $\mathcal{F}_{\alpha,\beta}$ denote the set of all such functions.

The above composite form of the objective function has been widely adopted in the literature on CM and CD (Wright,

2015). The separably non-smooth component ϕ arises naturally in many machine learning problems that often involve separable regularization such as ℓ_1 norm and box constraints.

2.2. Regret, Consistency, and Efficiency Measures

A stochastic optimization algorithm $\Upsilon = \{\Upsilon_t\}_{t=1}^T$ is a sequence of mappings from past actions and observations to the next choice of query point. The performance of Υ is measured by the cumulative regret defined as

$$R_\Upsilon(T) = \mathbb{E}\left[\sum_{t=1}^{T} F(\mathbf{x}_t, \xi_t) - F(\mathbf{x}^*, \xi_t)\right] \tag{6}$$

where the expectation is with respect to the random process of the query points and gradient observations induced by the algorithm Υ under the i.i.d. endogenous process of $\{\xi_t\}_{t=1}^T$.

In general, the performance of an algorithm depends on the underlying unknown objective function f (a dependency omitted in the regret notation for simplicity). Consider, for example, an algorithm that simply chooses one function in $\mathcal{F}_{\alpha,\beta}$ and sets its query points $\mathbf{x}_t$ to the minimizer of this function for all t would perform perfectly for the chosen function but suffers a linear regret order for all objective functions with sufficient deviation from the chosen one. It goes without saying that such heavily biased algorithms that completely forgo learning are of little interest.

We are interested in algorithms that offer good performance for all functions in $\mathcal{F}_{\alpha,\beta}$. An algorithm Υ is *consistent* if for all $f \in \mathcal{F}_{\alpha,\beta}$, the end point $\mathbf{x}_T$ produced by Υ satisfies

$$\lim_{T\to\infty} \mathbb{E}[f(\mathbf{x}_T)] = f(\mathbf{x}^*). \tag{7}$$

A consistent algorithm offers a sublinear regret order. This is also known as Hannan consistency or no-regret learning (Hannan, 1957). The latter term makes explicit the diminishing behavior of the average regret per action.

To measure the convergence rate of an algorithm, we introduce the concept of *p-consistency*. For a parameter $p \in (0, 1)$, we say Υ is *p-consistent* if

$$\sup_{f\in\mathcal{F}_{\alpha,\beta}} (\mathbb{E}[f(\mathbf{x}_T)] - f(\mathbf{x}^*)) \sim \Theta(T^{-p}). \tag{8}$$

A p-consistent algorithm offers an $O(T^{1-p})$ regret order for all $f \in \mathcal{F}_{\alpha,\beta}$. The parameter p measures the convergence rate.

An *efficient* algorithm is one that achieves the optimal convergence rate, hence lowest regret order. Specifically, Υ is *efficient* if for all initial query points $\mathbf{x}^{(1)} \in \mathcal{X}$, the end point $\mathbf{x}_T$ produced by Υ satisfies, for some $\lambda > 0$,

$$\sup_{f\in\mathcal{F}_{\alpha,\beta}} (\mathbb{E}[f(\mathbf{x}_T))] - f(\mathbf{x}^*)) \sim (f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*))^\lambda \Theta(T^{-1}). \tag{9}$$

An efficient algorithm offers the optimal $\log T$ regret order for all $f \in \mathcal{F}_{\alpha,\beta}$. In addition, it is able to leverage favorable initial conditions when they occur. We note here that the specific value of λ affects only the leading constant, but not the regret order. Hence for simplicity, we often use p-consistency with $p = 1$ to refer to efficient algorithms.

3. Progressive Coordinate Minimization

In this section, we present the PCM framework for extending low-dimensional optimization routines to high-dimensional problems. After specifying the general structure of PCM, we lay out the optimality criteria for designing its constituent components.

3.1. The General Structure of PCM

Within the PCM framework, an algorithm is given by a tuple $\Upsilon\,(\{\epsilon_k\}, \upsilon, \tau)$, where $\{\epsilon_k\}_{k\in\mathbb{N}}$ governs the progressive precision of each CM iteration indexed by k, υ is the low-dimensional optimization routine employed for coordinate minimization, and τ is the self-termination rule (i.e., a stopping time) for stopping υ at the given precision ϵ_k in each iteration k. Let $\tau(\epsilon)$ denote the (random) stopping time for achieving ϵ-precision under the termination rule τ.

A PCM algorithm $\Upsilon\,(\{\epsilon_k\}, \upsilon, \tau)$ operates as follows. At $t = 1$, an initial query point $\mathbf{x}^{(1)}$ and coordinate i_1 are chosen at random. The CM routine υ is then carried out along coordinate i_1 with all other coordinates fixed at $\mathbf{x}^{(1)}_{-i_1}$. At time $\tau(\epsilon_1)$, the first CM iteration ends and returns its last query point $\mathbf{x}_{\tau(\epsilon_1),\, i_1}$. The second iteration starts along a coordinate i_2 chosen uniformly at random and with the i_1 coordinate updated to its new value $\mathbf{x}_{\tau(\epsilon_1),\, i_1}$. The process repeats until the end of horizon T (see Algorithm 1 below).

3.2. Optimal Design of Constituent Components

PCM presents a framework for extending low-dimensional optimization algorithms to high-dimensional problems. The CM routine υ in a PCM algorithm is thus given, and we allow it to be an arbitrary p-consistent algorithm for any $p \in (0, 1]$ (note that the definitions of p-consistency and efficiency in Sec. 2 apply to arbitrary dimension.) Allowing arbitrary low-dimensional routines make PCM generally applicable, and the inclusion of consistent but not efficient (i.e., $p < 1$) routines responds to the shifted needs for low-cost solutions of only modest accuracy, as seen in modern machine learning and data analytics applications.

It is readily seen that for every $f(\mathbf{x}) \in \mathcal{F}_{\alpha,\beta}$, its low-

Algorithm 1 PCM $\Upsilon\left(\{\epsilon_k\}, \upsilon, \tau\right)$

Input: initial point $\mathbf{x}^{(1)}$.
Set $k \leftarrow 1$, $t \leftarrow 1$
repeat
  Choose coordinate i_k uniformly at random.
  Carry out υ along the direction i_k as follows:
    Set the initial point to $\mathbf{x}_{i_k}^{(k)}$ with fixed $\mathbf{x}_{-i_k}^{(k)}$.
    Continue until $\tau(\epsilon_k)$.
    Return the final point $\mathbf{x}_{\tau(\epsilon_k),i_k}$.
  $\mathbf{x}^{(k+1)} \leftarrow \left(\mathbf{x}_{\tau(\epsilon_k),i_k},\, \mathbf{x}_{-i_k}^{(k)}\right)$
  $k \leftarrow k+1$
  $t \leftarrow t + \tau(\epsilon_k)$
until $t = T$

dimensional restriction $f(\cdot, \mathbf{x}_{-i})$ for arbitrarily fixed $\mathbf{x}_{-i}$ belongs in $\mathcal{F}_{\alpha,\beta}$. Consequently, for a given low-dimensional routine υ with a certain consistency/efficiency level $p \in (0,1]$ (which needs to hold for all low-dimensional restrictions in $\mathcal{F}_{\alpha,\beta}$; see (8), (9)), its high-dimensional extension cannot have a better consistency level (or equivalently, lower regret order). The best possible outcome is that the high-dimensional extension preserves the p-consistency and the regret order of the low-dimensional algorithm for high-dimensional problems.

The design objective of PCM is thus to choose $\{\epsilon_k\}$ and τ for a given low-dimensional p-consistent algorithm υ such that the resulting high-dimensional algorithm preserves the regret order of υ.

The above optimization can be decoupled into two steps. First, the termination rule τ is designed to meet an order-optimal criterion as specified below. The optimal design of τ is specific to the routine υ, as one would expect. In the second step, the progression of precision $\{\epsilon_k\}$ is optimized for the given υ augmented with the order-optimal τ to preserve the p-consistency. Quite surprisingly, as shown in Sec. 4, there exists a *universal* optimal $\{\epsilon_k\}$ that is independent of not only the specific routine υ but also the specific consistency value $p \in (0,1]$.

Definition 1. *For a given p-consistent ($p \in (0,1]$) low-dimensional algorithm υ and a given $\epsilon > 0$, let $\tau(\epsilon)$ denote a stopping time over the random process of $\{x_t\}_{t\geq 1}$ induced by υ that satisfies $\mathbb{E}[f(x_{\tau(\epsilon)})] - f(x^*) \leq \epsilon$. A termination rule τ is order optimal in ϵ if*

$$\sup_{f\in\mathcal{F}_{\alpha,\beta}} \mathbb{E}[\tau(\epsilon)] \sim \Theta(\epsilon^{-1/p}). \tag{10}$$

Note that the above definition is for the dimensionality as determined by the given algorithm υ with f and $\mathcal{F}_{\alpha,\beta}$ defined accordingly.

An order-optimal termination rule is one that realizes the exponent p of the consistency of the underlying algorithm υ. The design of such termination rules is specific to υ, which we illustrate in Sec. 5 for two representative efficient (i.e., $p = 1$) low-dimensional routines.

4. The Optimal Precision Control

The theorem below establishes the optimal design of $\{\epsilon_k\}$ for arbitrary p-consistent low-dimensional routines.

Theorem 1. *Let υ be an arbitrary p-consistent ($p \in (0,1]$) low-dimensional routine and τ an order-optimal termination rule. For all $\gamma \in [(1-\alpha/(d\beta))^{1/2}, 1)$ and $\epsilon_0 > 0$, the PCM algorithm $\Upsilon\left(\{\epsilon_0\gamma^k\}, \upsilon, \tau\right)$ achieves a regret of $O(T^{1-p}\log^p T)$ for all $f \in \mathcal{F}_{\alpha,\beta}$.*

Theorem 1 shows that setting $\epsilon_k = \epsilon_0\gamma^k$ preserves the regret order[3] of υ. It is thus optimal. Such a choice of $\{\epsilon_k\}$ is independent of υ as well as the consistency level p of υ.

The proof of Theorem 1 is based on a decomposition of the regret as given below. Let K denote the (random) number of iterations until time T. Let $\mathbf{x}_{i_k}^* = \arg\min_x f((\mathbf{x}_{-i_k}^{(k-1)}, x))$ be the minimizer in the i_k^{th} coordinate with other coordinates fixed to $\mathbf{x}_{-i_k}^{(k-1)}$ (i.e., values from the previous iteration). Let $\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})} = (\mathbf{x}_{-i_k}^{(k-1)}, \mathbf{x}_{i_k}^*)$. Let $t_k = t_{k-1} + \tau_\upsilon(\epsilon_k)$ with $t_0 = 0$ denote the (random) time instants that mark the end of each iteration. We then have

$$\begin{aligned} R_\Upsilon(T) &= \mathbb{E}\left[\sum_{t=1}^{T} F(\mathbf{x}_t,\xi_t) - F(\mathbf{x}^*,\xi_t)\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_{k-1}+1}^{t_k} F(\mathbf{x}_t,\xi_t) - F(\mathbf{x}^*,\xi_t)\right]. \end{aligned}$$

This can be split into two terms using the local minimizer as

$$\begin{aligned} R_\Upsilon(T) = &\underbrace{\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_{k-1}+1}^{t_k}\left[F(\mathbf{x}_t,\xi_t) - F(\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})})\right]\right]}_{R_1} \\ &+ \underbrace{\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_{k-1}+1}^{t_k}\left[F(\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})}) - F(\mathbf{x}^*)\right]\right]}_{R_2}. \end{aligned} \tag{11}$$

The first term R_1 corresponds to the regret incurred by the low-dimensional routine υ carried out along one dimension.

[3] The preservation of the regret order is exact for efficient routines. For consistent but not efficient (i.e., $p < 1$) routines, the preservation is up to a poly-log term which is dominated by the term of T^{1-p}.

Note that this regret is computed with respect to the one-dimensional local minima $\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})}$. The second term R_2 corresponds to the loss incurred at the one-dimensional local minima in excess to the global minimum $\mathbf{x}^*$.

The above regret decomposition also provides insight into the optimal design of $\{\epsilon_k\}$. To achieve a low regret order, it is desirable to equalize the orders of R_1 and R_2. If a more aggressive choice of ϵ_k is used, then the rate of decay of the CM iterates is unable to compensate for the time required for higher accuracy, resulting in R_2 dominating R_1. On the other hand, a more conservative choice will lead to a slower decay in objective function with an increased number of CM iterations. This would result in increasing both the terms to an extent where Υ will no longer be able to maintain the consistency level of υ.

Proof. We give here a sketch of the proof. The analysis of R_1 and R_2 builds on analytical characterizations of the following two key quantities: the expected number $\mathbb{E}[K]$ of CM iterations and the convergence rate of CM outputs $\{\mathbf{x}^{(k)}\}_{k=1}^K$. They are given in the following two lemmas.

Lemma 1. *Let υ be a p-consistent policy for some $p \in (0,1]$ and τ its order-optimal termination rule. Under $\Upsilon\left(\{\epsilon_0\gamma^k\}, \upsilon, \tau\right)$, we have $\mathbb{E}[K] \sim O(\log T)$ for all $f \in \mathcal{F}_{\alpha,\beta}$.*

Lemma 2. *Let $\{\mathbf{x}^{(k)}\}$ be the sequence of CM outputs generated under $\Upsilon\left(\{\epsilon_0\gamma^k\}, \upsilon, \tau\right)$ for a function $f \in \mathcal{F}_{\alpha,\beta}$. Then the sequence of points $\{\mathbf{x}^{(k)}\}$ satisfy $\mathbb{E}[f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)] \leq F_0\gamma^k$ for all $k \geq 0$ and for all $\gamma \in [(1-\alpha/(d\beta))^{1/2}, 1)$ where $F_0 = \max\{f(\mathbf{x}_0) - f(\mathbf{x}^*), \epsilon_0/(1-\gamma)\}$.*

R_1 is bounded using the consistency level of the routine υ augmented with the termination rule τ. R_2 is bounded using Lemma 2 and the expected time taken in each CM iteration. On plugging in the value of ϵ_k, both terms end up being of the same order and we arrive at the theorem. The detailed proofs of the lemmas and the theorem can be found in the supplementary material. □

5. Termination Rules

In this section, we illustrate the construction of order-optimal termination rules for two representative and fundamentally different low-dimensional routines, one classical, one recent. For simplicity, we focus on smooth objective functions. All notations are for a specific coordinate with coordinate index omitted.

5.1. SGD

For a given initial point x_1, SGD proceeds by generating the following sequence of query points

$$x_{t+1} = \text{proj}_{\mathcal{X}}(x_t - \eta_t G(x_t, \xi_t)), \tag{12}$$

where $G(x_t, \xi_t)$ is the random gradient observed at x_t, $\{\eta_t\}_{t\geq 1}$ is the sequence of step sizes, and $\text{proj}_{\mathcal{X}}$ denotes the projection operator onto the convex set $\mathcal{X}$ (restricted to the chosen coordinate with other coordinates fixed). The following lemma establishes the efficiency of the SGD routine with properly chosen hyperparameters and the order optimality of the termination rule for noise models with bounded variance. Based on Theorem 1, we can then conclude that the resulting PCM algorithm with $\{\epsilon_k\} = \epsilon_0\gamma^k$ is an efficient algorithm with a regret of $O(\log T)$.

Lemma 3. *Consider the low-dimensional routine of SGD with step sizes given by $\eta_t = \mu/(1+\nu t)$ with $\mu = \dfrac{\mu_0\alpha}{2g_{\max}^2}$ and $\nu = \dfrac{\mu_0\alpha^2}{4g_{\max}^2}$, where $g_{\max}$ is an upper bound on the second moment of the random gradient, $G(x,\xi)$, for all $x \in \mathcal{X}$ and μ_0 a properly chosen hyperparameter. Then SGD with the above chosen parameters is an efficient algorithm as defined in (9). The termination rule given by $\tau(\epsilon) = \left\lceil \dfrac{2\beta g_{\max}^2}{\alpha^2\epsilon} \right\rceil$ is order optimal as defined in Definition 1.*

Proof. We give here a sketch of the proof. Details can be found in the supplementary material. The order-optimality of the termination rule follows immediately from definition 1. For implementation in PCM, the constant μ_0 is set to $\mu_0 \sim \Theta(\gamma^k)$ in iteration k to ensure the adaptivity to the initial point. Using smoothness of f, we obtain $\mathbb{E}[f(x_t) - f(x^*)] \leq \beta\mathbb{E}[|x_t - x^*|^2] \leq \mu_0/(1+\nu t)$, implying that SGD is a consistent algorithm with $p = 1$. The choice of μ_0 makes it an efficient algorithm with $\lambda = 1$. □

5.2. RWT

RWT (Random Walk on a Tree) proposed in (Vakili & Zhao, 2019) is restricted to one-dimensional problems. There do not appear to be any simple extensions of RWT to high-dimensional problems. We show here that PCM offers a possibility and preserves its efficiency.

Without loss of generality, assume that the one-dimensional domain is the closed interval $[0,1]$. The basic idea of RWT is to construct an infinite-depth binary tree based on successive partitions of the interval. Each node of the tree represents a sub-interval, and nodes on the same level give an equal-length partition of $[0,1]$. The query point at each time is then generated based on a biased random walk on the interval tree that initiates at the root and is biased toward the infinitesimally small interval containing the minimizer x^*.

When the random walk reaches a node, the two end points along with the middle point of the corresponding interval are queried in serial to determine, with a required confidence level $\breve{p}$, the sign of $g(x)$ at those points. The test on the sign of $g(x)$ at any given x is done through a confidence-bound based local sequential test using random gradient observations. The outcomes of the sign tests at the three points of the interval determines the next move of the random walk: to the child that contains a sign change or back to the parent under inconsistent test outcomes.The confidence level $\breve{p}$ of the sign test ensures the bias of the walk, i.e., the probability of moving toward x^* is greater than $1/2$ at each step. For one-dimensional problems, the biased walk on the tree continues until T.

To extend RWT to high-dimensional problems within the PCM framework, we propose the following termination rule. Specifically, we leverage the structure of the local confidence-bound based sequential test in RWT. Note that the precision condition required at termination can be translated to an upper bound on the magnitude of the gradient. Since the local test is designed to estimate the sign of the gradient, it naturally requires more samples as the gradient gets closer to zero (i.e., the signal strength reduces while the noise persists). Assume that the noise is sub-Gaussian, that is, the moment generating function of $G_i(\mathbf{x};\xi) - g_i(\mathbf{x})$ is upper bounded by that of a Gaussian with variance σ_i^2 for $i = 1, 2, \ldots, d$. We propose the following termination rule: the current CM iteration terminates once a sequential test draws more than $N_0(\epsilon) = \dfrac{40\sigma_0^2}{\alpha\epsilon} \log\left(\dfrac{2}{\breve{p}} \log\left(\dfrac{80\sigma_0^2}{\alpha\breve{p}\epsilon}\right)\right)$ samples, where $\sigma_0^2 \geq \max_i \sigma_i^2$.

This threshold is so designed that when the number of samples in the sequential test exceeds that value, the gradient at that point is sufficiently small with high probability, leading to the required precision. It is interesting to note that the termination rule for SGD given in Lemma 3 is an open-loop design with pre-fixed termination time, while the termination rule proposed for RWT is closed-loop and adapts to random observations.

The following lemma gives the regret order of the high-dimensional extension of RWT within the PCM framework. The detailed proof of the lemma is given in the supplementary material.

Lemma 4. *The PCM-RWT algorithm with the chosen termination rule has a regret order of* $O(\log T(\log\log T)^2)$.

6. Discussions and Extensions

Parallel and Distributed Computing: One of the major advantages of CD/CM based methods is their amenability to parallel and distributed computing, which PCM naturally inherits. Advantages of CD/CM methods in parallel and distributed computing have been well studied in the literature (Richtárik & Takáč, 2016a;b; Bradley et al., 2011; Peng et al., 2013; Liu et al., 2015; Mareček et al., 2015; Richtárik & Takáč, 2016). It has been shown that the convergence of coordinate-based methods with parallel updates is guaranteed only when the parallel updates are aggregated in such a way that the combined update leads to a decrease in the function value as compared to the previous iterate. Such a condition is possible to verify and enforce when the objective function is deterministic and known, but presents a challenge in stochastic optimization.

To achieve parallelization of PCM that maintains the p-consistency as in the serial implementation, we draw inspiration from the technique proposed in (Ferris & Mangasarian, 1994) that leverages the convexity of the objective function.

Assume there are $m < d$ independent cores, connected in parallel to a main server. The current iterate is passed to all the cores, with each receiving a different coordinate index, chosen at random. After the one dimensional optimization completes at each core, the next query point is set to the average of points returned by all the cores. A decrease in the function value at the averaged point compared to the initial point is guaranteed by the convexity of the function. It can be shown that with a parallel implementation of PCM, the "effective" dimension of the problem is reduced from d to d/m (see details in supplementary material).

Heavy-Tailed Noise: The noise in the partial gradient estimates affects only the consistency level p of the low-dimensional routine υ, subsequently, the design of the order-optimal termination rule. The optimal precision control, however, is independent of the noise characteristics. More specifically, the same optimal precision control as specified in Theorem 1 preserves the p-consistency and regret order of the low-dimensional routine υ even under heavy-tailed noise. In particular, for heavy-tailed noise with a finite b^{th} moment ($b \in (1, 2)$), both SGD and RWT are $(2 - 2/b)$-consistent and offer an optimal regret order of $O(T^{2/b-1})$ (up to poly-$\log T$ factors) (Zhang et al., 2019; Vakili et al., 2019). It is not difficult to show that under heavy-tailed noise, an open-loop termination rule with $\tau(\epsilon) = C\epsilon^{\frac{b}{2(1-b)}}$ for SGD and a similar sample-threshold based termination rule with $N_0(\epsilon) = C'\epsilon^{\frac{b}{2(1-b)}}\,\text{polylog}(1/\epsilon)$ for RWT are order optimal, where C and C' are constants. In conclusion, PCM offers an optimal high-dimensional extension for any p-consistent algorithm regardless of the noise characteristics.

Zeroth-order feedback model: The two example low-dimensional stochastic optimization routines used in Sec. 5 are first-order algorithms that use gradient to update the query points. However, the PCM framework is equally applicable to zeroth-order algorithms which directly learn from

random losses $F(x_t;\xi)$. As it can be seen from Theorem 1, the PCM framework is agnostic to the low-dimensional routine υ and the associated termination rule (provided it is order optimal).

7. Empirical Results

In this section, we compare the regret performance of PCM employing SGD with that of SCD (based on its online version developed in (Wang & Banerjee, 2014)). We consider the problem of binary classification on the MNIST dataset (Lecun et al., 1998) as a one-vs-rest classification problem in an online setting. We use regularized hinge loss as the loss function. At each time t, $\xi_t = (Y_t, Z_t)$ is drawn uniformly at random from the dataset. Then using the current classifier $\mathbf{x}_t$, we incur a random loss

$$F(\mathbf{x}_t, \xi_t) = \max\{0, 1 - Z_t \langle \mathbf{x}_t, Y_t \rangle\} + \frac{\alpha}{2}\|\mathbf{x}_t\|^2 \quad (13)$$

and observe the partial gradient given as

$$G_{i_t}(\mathbf{x}_t, \xi_t) = -Z_t (Y_t)_{i_t} \mathbb{1}\{1 - Z_t \langle \mathbf{x}_t, Y_t \rangle > 0\} + \alpha(\mathbf{x}_t)_{i_t} \quad (14)$$

where i_t denotes the index of the coordinate chosen at time t. Both algorithms are randomly initialised with a point in $[-0.5, 0.5]^d$ and run for $T = 1000d$, where $d = 785$ is the dimensionality of the problem. The regularization parameter is set to $\alpha = 1.2 \times 10^{-2}$. The regret plotted is averaged over 10 Monte Carlo runs. The SCD algorithm is run with stepsize $\eta_t = 5/\lceil t/10000 \rceil$. The PCM algorithm is implemented using a SGD as the local optimization routine with $\gamma = 0.99999$, $\epsilon_0 = 0.1$ and step size of $\eta = 0.2$. The step size in each iteration was reduced by a factor of γ. The termination rule was set to $\lceil 1/2\epsilon_k \rceil$. The parameters in both algorithms were optimized based on a grid search. The results obtained for various digits are shown in Figure 1. It is evident from the plots in Figure 1 that PCM has a significantly better performance. It suggests that the advantages of CM over CD type methods also hold in stochastic optimization.

8. Conclusion

We considered the problem of stochastic convex optimization where an unknown stochastic loss function is minimized based on noisy estimates of partial (sub-)gradients. We developed a stochastic coordinate minimization framework that extends any given low-dimensional stochastic optimization routine to high-dimensional problems and preserves its regret order. The crux of our approach is an optimal control of the progressive precision of the CM iterations. For strongly-convex functions, we established a universally optimal precision sequence that is agnostic to the low-dimensional routine and its consistency level, as well as the noise characteristics and feedback model of the underlying high-dimensional problem.

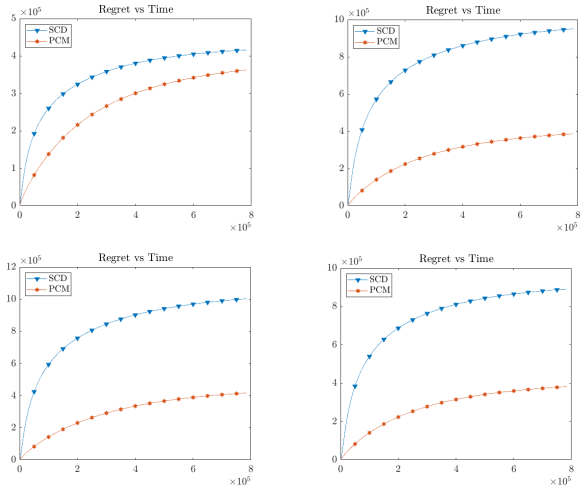

Figure 1. From top left, in clockwise order, the digits are $1, 2, 4, 3$. We obtain similar results for all the other digits as well.

We have focused in this work on strongly convex and separably non-smooth functions. While the PCM framework applies to general objective functions, the optimal design of the precision sequence and the resulting regret performance remain open. Investigation in this direction is currently underway.

9. Acknowledgements

The work of Sudeep Salgia and Qing Zhao was supported by the National Science Foundation under Grant CCF-1815559. We would also like to thank the anonymous reviewers for their constructive comments.

References

Beck, A. and Tetruashvili, L. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013. ISSN 10526234. doi: 10.1137/120887679.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. ISSN 00361445. doi: 10.1137/16M1080173.

Bradley, J. K., Kyrola, A., Bickson, D., and Guestrin, C. Parallel coordinate descent for L1-regularized loss minimization. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 321–328, 2011. ISBN 9781450306195.

Csiba, D., Qu, Z., and Richtarik, P. Stochastic dual coor-

dinate ascent with adaptive probabilities. In *32nd International Conference on Machine Learning, ICML 2015*, volume 1, pp. 674–683, 2015. ISBN 9781510810587.

Dang, C. D. and Lan, G. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization*, 25(2):856–881, 2015. ISSN 10526234. doi: 10.1137/130936361.

Deng, Q., Ho, J., and Rangarajan, A. Stochastic coordinate descent for nonsmooth convex optimization. 12 2013.

Fercoq, O. and Richtárik, P. Smooth Minimization of Nonsmooth Functions with Parallel Coordinate Descent Methods. In *Springer Proceedings in Mathematics and Statistics*, volume 279, pp. 57–96, 2019. ISBN 9783030121181. doi: 10.1007/978-3-030-12119-8_4.

Ferris, M. C. and Mangasarian, O. L. Parallel Variable Distribution. *SIAM Journal on Optimization*, 4(4):815–832, 1994. ISSN 1052-6234. doi: 10.1137/0804047.

Grippo, L. and Sciandrone, M. Globally convergent block-coordinate techniques for unconstrained optimization. *Optimization Methods and Software*, 10(4): 587–637, 1999. ISSN 10556788. doi: 10.1080/10556789908805730.

Hannan, J. Approximation to rayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

Hsieh, C.-J., Chang, K.-W., Lin, C.-J., Keerthi, S. S., and Sundararajan, S. Coordinate Descent Method for Large-scale L2-loss Linear Support Vector Machines. *Journal of Machine Learning Research*, 9:1369–1398, 2008a.

Hsieh, C. J., Chang, K. W., Lin, C. J., Keerthi, S. S., and Sundararajan, S. A dual coordinate descent method for large-scale linear SVM. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 408–415, 2008b. ISBN 9781605582054.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9851 LNAI, pp. 795–811, 2016. ISBN 9783319461274. doi: 10.1007/978-3-319-46128-1_50.

Konečný, J., Liu, J., Richtárik, P., and Takáč, M. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10 (2):242–255, March 2016. ISSN 1941-0484. doi: 10.1109/JSTSP.2015.2505682.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. ISSN 1558-2256. doi: 10.1109/5.726791.

Leventhal, D. and Lewis, A. S. Randomized methods for linear constraints: Convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010. ISSN 0364765X. doi: 10.1287/moor.1100.0456.

Liu, J., Wright, S. J., Ré, C., and Sridhar, S. An Asynchronous Parallel Stochastic Coordinate Descent Algorithm. *Journal of Machine Learning Research*, 16:285–322, 2015.

Lu, Z. and Xiao, L. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, 2015. ISSN 14364646. doi: 10.1007/s10107-014-0800-2.

Luo, Z. Q. and Tseng, P. On the Convergence of the Coordinate Descent Method for Convex Differentiable Minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

Mareček, J., Richtárik, P., and Takáč, M. Distributed block coordinate descent for minimizing partially separable functions. In *Springer Proceedings in Mathematics and Statistics*, volume 134, pp. 261–288, 2015. ISBN 9783319176888. doi: 10.1007/978-3-319-17689-5_11.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. ISSN 10526234. doi: 10.1137/100802001.

Nesterov, Y. Subgradient methods for huge-scale optimization problems. *Mathematical Programming*, 146 (1-2):275–297, 2014. ISSN 14364646. doi: 10.1007/s10107-013-0686-4.

Pasupathy, R., Tech, V., and Kim, S. The Stochastic Root Finding Problem: Overview, Solutions, and Open Questions. *ACM Transactions on Modeling and Computational Simulations*, 21(3):19, 2011.

Peng, Z., Yan, M., and Yin, W. Parallel and distributed sparse optimization. In *Asilomar Conference on Signals, Systems and Computers*, pp. 659–664, 2013. ISBN 9781479923908. doi: 10.1109/ACSSC.2013.6810364.

Razaviyayn, M., Hong, M., and Luo, Z. Q. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013. ISSN 10526234. doi: 10.1137/120891009.

Reddi, S. J., Hefny, A., Downey, C., Dubey, A., and Sra, S. Large-scale randomized-coordinate descent methods with non-separable linear constraints. In *Uncertainty in Artificial Intelligence - Proceedings of the 31st Conference, UAI 2015*, pp. 762–771, 2015.

Richtárik, P. and Takáč, M. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014. ISSN 14364646. doi: 10.1007/s10107-012-0614-z.

Richtárik, P. and Takáč, M. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, 2016a. ISSN 14364646. doi: 10.1007/s10107-015-0901-6.

Richtárik, P. and Takáč, M. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17:1–25, 2016b. ISSN 15337928.

Richtárik, P. and Takáč, M. On optimal probabilities in stochastic coordinate descent methods. *Optimization Letters*, 10(6):1233–1243, Aug 2016. ISSN 1862-4480. doi: 10.1007/s11590-015-0916-1.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Statistics*, 22(3):400–407, 1951.

Ruder, S. An overview of gradient descent optimization algorithms. *ArXiv*, abs/1609.04747, 2016.

Saha, A. and Tewari, A. On the finite time convergence of cyclic coordinate descent methods. *SIAM Journal on Optimization*, 23(1):576–601, 2013. ISSN 10526234. doi: 10.1137/110840054.

Salehi, F., Thiran, P., and Elisa Celis, L. Coordinate descent with bandit sampling. In *Advances in Neural Information Processing Systems*, pp. 9247–9257. Curran Associates, Inc., 2018.

Shalev-Shwartz, S. and Zhang, T. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 64–72, Beijing, China, 22–24 Jun 2014. PMLR.

Tao, Q., Kong, K., Chu, D., and Wu, G. Stochastic coordinate descent methods for regularized smooth and nonsmooth losses. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7523, pp. 537–552, 2012. ISBN 9783642334597. doi: 10.1007/978-3-642-33460-3_40.

Tappenden, R., Richtárik, P., and Gondzio, J. Inexact Coordinate Descent: Complexity and Preconditioning. *Journal of Optimization Theory and Applications*, 170 (1):144–176, 2016. ISSN 15732878. doi: 10.1007/s10957-016-0867-4.

Tewari, A. and Shalev-Shwartz, S. Stochastic Methods for l1-regularized Loss Minimization. *Journal of Machine Learning Research*, 12:1865–1892, 2011.

Tibshirani, R. Coordinate descent. pp. 1–28, 2013. URL https://www.stat.cmu.edu/~ryantibs/convexopt/lectures/coord-desc.pdf.

Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001. ISSN 00223239. doi: 10.1023/A:1017501703105.

Tseng, P. and Yun, S. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *Journal of Optimization Theory and Applications*, 140(3):513, Sep 2008. ISSN 1573-2878. doi: 10.1007/s10957-008-9458-3.

Tseng, P. and Yun, S. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009. ISSN 00255610. doi: 10.1007/s10107-007-0170-0.

Vakili, S. and Zhao, Q. A random walk approach to first-order stochastic convex optimization. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 395–399, July 2019. doi: 10.1109/ISIT.2019.8849396.

Vakili, S., Salgia, S., and Zhao, Q. Stochastic Gradient Descent on a Tree: An Adaptive and Robust Approach to Stochastic Convex Optimization. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2019*, pp. 432–438, 2019. ISBN 9781728131511. doi: 10.1109/ALLERTON.2019.8919740.

Wang, H. and Banerjee, A. Randomized Block Coordinate Descent for Online and Stochastic Optimization. 2014.

Wright, S. J. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015. ISSN 14364646. doi: 10.1007/s10107-015-0892-3.

Xu, Y. and Yin, W. Block stochastic gradient iteration for convex and nonconvex optimization. *SIAM Journal on Optimization*, 25(3):1686–1716, 2015. ISSN 10526234. doi: 10.1137/140983938.

Zhang, A. and Gu, Q. Accelerated stochastic block coordinate descent with optimal sampling. In *Proceedings of the ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining, volume 13-17-Augu, pp. 2035–2044, 2016. ISBN 9781450342322. doi: 10.1145/2939672.2939819.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S. J., Kumar, S., and Sra, S. Why ADAM Beats SGD for Attention Models. 2019.

Zhao, T., Yu, M., Wang, Y., Arora, R., and Liu, H. Accelerated mini-batch randomized block coordinate descent method. In *Advances in Neural Information Processing Systems*, volume 4, pp. 3329–3337, 2014.