Linear Convergence of Randomized Primal-Dual Coordinate Method for Large-scale Linear Constrained Convex Programming

Daoli Zhu [*1] **Lei Zhao** [*2]

Abstract

Linear constrained convex programming has many practical applications, including support vector machine and machine learning portfolio problems. We propose the randomized primal-dual coordinate (RPDC) method, a randomized coordinate extension of the first-order primal-dual method by (Cohen & Zhu, 1984) and (Zhao & Zhu, 2019), to solve linear constrained convex programming. We randomly choose a block of variables based on a uniform distribution, linearize, and apply a Bregman-like function (core function) to the selected block to obtain simple parallel primal-dual decomposition. We then establish almost surely convergence and expected $O(1/t)$ convergence rate, and expected linear convergence under global strong metric subregularity. Finally, we discuss implementation details for the randomized primal-dual coordinate approach and present numerical experiments on support vector machine and machine learning portfolio problems to verify the linear convergence.

1. Introduction

This paper considers linear constrained convex programming (LCCP),

$$\text{(P):}\quad \begin{array}{ll} \min & F(u) = G(u) + J(u) \\ \text{s.t} & Au - b = 0 \\ & u \in \mathbf{U} \end{array}, \tag{1}$$

where G is a convex smooth function on the closed convex set $\mathbf{U} \subset \mathbf{R}^n$; and J is a convex, possibly non-smooth function on $\mathbf{U} \subset \mathbf{R}^n$. We assume that $J(u) = \sum_{i=1}^{N} J_i(u_i)$ is additive with respect to the space decomposition

$$\mathbf{U} = \mathbf{U}_1 \times \mathbf{U}_2 \cdots \times \mathbf{U}_N, u_i \in \mathbf{U}_i \subset \mathbf{R}^{n_i} \text{ and } \sum_{i=1}^{N} n_i = n. \tag{2}$$

Each J_i is a convex but possibly non-smooth function on $U_i \subset \mathbf{R}^{n_i}$; $A = (A_1, A_2, \cdots, A_N) \in \mathbf{R}^{m\times n}$ is an appropriate partition of A, where A_i is an $m \times n_i$ matrix and $b \in \mathbf{R}^m$ is a vector.

1.1. Motivation

Linear constrained convex programming is an important and challenging application problem class. We present several example applications to demonstrate the reasons for interest in type (P) problems.

1.1.1. Support vector machine

Support vector machine (SVM) is a popular supervised learning method (Boser et al., 1992; Cortes & Vapnik, 1995), widely used for pattern recognition (Burges, 1998; Schölkopf et al., 2000) and classification (Chang & Lin, 2011). The SVM problem can be expressed as

$$\text{(SVM)}\quad \begin{array}{ll} \min\limits_{u\in[0,c]^n} & \frac{1}{2}u^\top Qu - \mathbf{1}_n^\top u \\ \text{s.t.} & y^\top u = 0 \end{array},$$

where $u \in \mathbf{R}^n$ are the decision variables, $Q \in \mathbf{R}^{n\times n}$ is a symmetric and positive-definite matrix, $c \in \mathbf{R}$ is the upperbound of all variables, $y \in \{-1, 1\}^n$ is the vector of labels, and $\mathbf{1}_n$ is an n-dimensional vector of 1s.

1.1.2. Machine learning portfolio problem

Portfolio optimization (PO) via machine learning has received increased attention recently. PO aims to invest in a group of financial assets with instructions by machine, based on financial principles and optimization strategies. Since this requires considerable quantitative calculation, machine learning methods are essential to reduce human mistakes and biases for real-world investment. (Brodie et al., 2009; Lai et al., 2018; Li et al., 2016; Ho et al., 2015; Shen et al., 2014) The machine learning portfolio (MLP) problem can

[*]Equal contribution [1]Antai College of Economics and Management and Sino-US Global Logistics Institute, Shanghai Jiao Tong University, Shanghai, China [2]School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai, China. Correspondence to: Lei Zhao <l.zhao@sjtu.edu.cn>.

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

be expressed as

$$\text{(MLP)} \quad \begin{array}{ll} \min\limits_{u\in\mathbf{R}^n} & \frac{1}{2}u^\top \Sigma u + \lambda\|u\|_1 \\ \text{s.t.} & \mu^\top u = \rho \\ & \mathbf{1}_n^\top u = 1 \end{array},$$

where $u \in \mathbf{R}^n$ is the decision portfolio vector; $\Sigma \in \mathbf{R}^{n\times n}$, which is symmetric and positive-definite, is the estimated covariance matrix of asset returns; $\mu \in \mathbf{R}^n$ is the expectation of asset returns; ρ is a predefined prospective growth rate; and $\mathbf{1}_n$ is an n-dimensional vector of 1s.

1.2. Related works

The augmented Lagrangian method (ALM) is an approach for general LCCP that can overcome the dual Lagrangian instability and non-differentiability. The augmented Lagrangian for a constrained convex program has the same solution set as the original constrained convex program. Consider the following augmented Lagrangian function of (P),

$$L_\gamma(u,p) = F(u) + \langle p, Au - b\rangle + \frac{\gamma}{2}\|Au-b\|^2. \tag{3}$$

The ALM for equality-constrained optimization problems can be expressed as (Hestenes, 1969; Powell, 1969)

ALM

$$\begin{cases} u^{k+1} = \arg\min\limits_{u\in\mathbf{U}} L_\gamma(u,p^k); \\ p^{k+1} = p^k + \gamma(Au^{k+1}-b). \end{cases}$$

Although ALM has several advantages, it does not preserve separability, even when the initial problem is separable. One way to decompose the augmented Lagrangian is to use alternating direction method of multipliers (ADMM) (Fortin & Glowinski, 1983), which applies a Gauss-Seidel-like minimization strategy. Another way to overcome this difficulty is the auxiliary problem principle of augmented Lagrangian (APP-AL) (Cohen & Zhu, 1984) method, a fairly general first-order primal-dual decomposition method based on linearizing the augmented Lagrangian of LCCP. The APP-AL scheme for LCCP can be expressed as

APP-AL

$$\begin{cases} u^{k+1} = \arg\min\limits_{u\in\mathbf{U}} \langle \nabla G(u^k), u\rangle + J(u) + \langle q^k, Au\rangle \\ \qquad\qquad\qquad\qquad\qquad\qquad + \frac{1}{\epsilon}D(u,u^k); \\ p^{k+1} = p^k + \gamma(Au^{k+1}-b), \end{cases}$$

where $q^k = p^k + \gamma(Au^k - b)$, $D(u,v) = K(u) - K(v) - \langle \nabla K(v), u - v\rangle$ is a Bregman like function with K is strongly convex and gradient Lipschitz. (Zhao & Zhu, 2019) extended (Cohen & Zhu, 1984) to propose first-order primal-dual augmented Lagrangian methods for nonlinear convex cone programming with separable and non-separable objective and constraints as an algorithm (variant auxiliary problem principle, VAPP). They showed that APP-AL can be viewed as a forward-backward splitting method to find a solution for (P): $w^{k+1} = (\Gamma^k + \mathcal{A})^{-1}(\Gamma^k - \mathcal{B})w^k$, where

$$\mathcal{A}(w) = \begin{pmatrix} \partial J(u) + \mathcal{N}_{\mathbf{U}}(u) \\ \mathbf{0}_m \end{pmatrix}$$

and

$$\mathcal{B}(w) = \begin{pmatrix} \nabla G(u) + A^\top[p + \gamma(Au-b)] \\ b - Au \end{pmatrix}$$

are two maximal monotone mappings; $w = (u,p)$; $\mathcal{N}_{\mathbf{U}}(u) = \{\xi : \langle \xi, \zeta - u\rangle \le 0, \forall\zeta \in \mathbf{U}\}$ is the normal cone at u to $\mathbf{U}$; and $\Gamma^k(w) = \begin{pmatrix} \frac{1}{\epsilon}\nabla K(u) \\ \frac{1}{\gamma}[p-p^k] - [Au-b] \end{pmatrix}$ is a strongly monotone nonlinear Bregman operator. Problem (P) can be reformulated as an augmented Lagrangian based inclusion problem,

$$\begin{aligned} 0 \in H_\gamma(w) &= \begin{pmatrix} \partial_u L_\gamma(u,p) + \mathcal{N}_{\mathbf{U}}(u) \\ \nabla_p L_\gamma(u,p) \end{pmatrix} \\ &= \begin{pmatrix} \nabla G(u) + \partial J(u) + A^\top[p+\gamma(Au-b)] \\ \qquad\qquad\qquad\qquad + \mathcal{N}_{\mathbf{U}}(u) \\ b - Au \end{pmatrix} \\ &= \mathcal{A}(w) + \mathcal{B}(w). \end{aligned} \tag{4}$$

We propose an iteration based forward-backward splitting algorithm to solve (4), $w^{k+1} = (\Gamma^k + \mathcal{A})^{-1}(\Gamma^k - \mathcal{B})w^k$, hence $(\Gamma^k - \mathcal{B})w^k \in (\Gamma^k + \mathcal{A})w^{k+1}$. Thus,

$$0 \in \begin{pmatrix} \frac{1}{\epsilon}[\nabla K(u^{k+1}) - \nabla K(u^k)] + \nabla G(u^k) \\ \qquad + A^\top[p^k + \gamma(Au^k - b)] + \partial J(u^{k+1}) \\ \qquad\qquad\qquad\qquad + \mathcal{N}_{\mathbf{U}}(u^{k+1}) \\ p^{k+1} - p^k - \gamma(Au^{k+1} - b) \end{pmatrix},$$

which is the exact inclusion problem formulation for APP-AL. Zhao and Zhu also proposed linear convergence for VAPP (Zhao & Zhu, 2019).

Recent big data applications employ very large datasets that are commonly distributed over different locations. Hence it is often impractical to assume that optimization algorithms can traverse an entire dataset once in each iteration, because this would be very time consuming and/or unreliable, often resulting in low resource utilization due to synchronization among different computing units, e.g. CPUs, GPUs, and cores, in a distributed computing environment. On the other hand, block coordinate descent (BCD) algorithms can make progress by using information obtained from a randomly selected data subsets and, hence provide more flexible implementation in distributed environments. The main advantage of BCD is to reduce complexity and memory requirements per iteration, which becomes increasingly important for very-large scale problems.

We briefly review some related works on coordinate type methods.

Two BCD variations are widely employed for problems without constraints. The first BCD variation relates to block-choosing strategy. One common approach is a cyclic strategy. (Tseng, 2001) proved BCD cyclic strategy convergence and (Luo & Tseng, 1992) and (Wang & Lin, 2014) proved local and global linear convergence, respectively, under specific assumptions. Another common approach is a randomized strategy. (Nesterov, 2012) studied the convergence rate of randomized BCD for convex smooth optimization and (Richtárik & Takáč, 2014) and (Lu & Xiao, 2015) subsequently extended Nesterov's technique to composite optimization. The second BCD variation relates to the point read to evaluate the gradient in each iteration. The approaches are called asynchronous BCD if the read points have different "ages", and synchronous BCD otherwise. All BCD variants reviewed above are synchronous BCD. (Liu & Wright, 2015) and (Liu et al., 2014) established the convergence rate of asynchronous BCD for composite optimization and convex smooth optimization, respectively, without constraints.

Few previous studies considered BCD methods for problems with constraints. (Necoara & Patrascu, 2014) proposed a random coordinate descent algorithm for an optimization problem with one linear constraint. (Gao et al., 2019) and (Xu & Zhang, 2018) considered a similar scheme to RPDC, obtaining expected $O(1/t)$ and $O(1/t^2)$ rates, respectively. (Xu, 2019) recently proposed an asynchronous RPDC algorithm with expected $O(1/t)$ rate. However, to the best of our knowledge, no previous study considered convergence and linear convergence results for RPDC.

This paper focused on RPDC, the randomized coordinate extension of APP-AL, as shown in Algorithm 1.

1.3. Main contributions and outline for this paper

We propose the randomized primal-dual coordinate (RPDC) method based on the first-order primal-dual method (Cohen & Zhu, 1984; Zhao & Zhu, 2019). The RPDC method randomly updates one block of variables based on a uniform distribution. The main contributions from this paper are as follows:

(i) We show that the sequence generated by RPDC converges to an optimal solution with probability 1.

(ii) We show RPDC has expected $O(1/t)$ rate for general LCCP.

(iii) We establish the expected linear convergence of RPDC under global strong metric subregularity.

(iv) We show that SVM and MLP problems satisfy global strong metric subregularity under some reasonable conditions.

(v) Finally, we discuss the implementation details of RPDC and present numerical experiments on SVM and MLP problems to verify linear convergence.

The remainder of this paper is organized as follows. Section 2 discusses technical preliminaries. Section 3 shows almost surely convergence and expected $O(1/t)$ convergence rate for RPDC. Section 4 establishes the expected linear convergence of RPDC under global strong metric subregularity. Section 5 discusses implementation details for RPDC and presents numerical experiments on SVM and MLP problems. Finally, Section 6 summarizes and concludes the paper.

2. Preliminaries

This section provides some useful preliminaries for subsequent discussions and summarizes notations and assumptions. We denote vector inner product and Euclidean norm as $\langle\cdot\rangle$ and $\|\cdot\|$, respectively.

2.1. Notations and assumptions

Throughout this paper, we make the following standard assumptions for Problem (P).

Assumption 1 (H_1) *J is a convex, lower semi-continuous function (not necessarily differentiable) such that $\mathbf{dom}J \cap \mathbf{U} \neq \emptyset$.*
(H_2) *G is convex and differentiable, and its derivative is Lipschitz with constant B_G.*
(H_3) *There exists at least one saddle point for the Lagrangian of* (P).

From Assumption 1 and Theorem 3.2.12 (Ortega & Rheinboldt, 1970), the following descent property for G holds

$$G(v) - G(u) - \langle \nabla G(u), v-u\rangle \leq \frac{B_G}{2}\|u-v\|^2. \tag{5}$$

2.2. Lagrangian and Karush-Kuhn-Tucker mapping

The Lagrangian of (P) is defined as

$$L(u,p) = F(u) + \langle p, Au-b\rangle, \tag{6}$$

and a saddle point $(u^*,p^*) \in \mathbf{U}\times\mathbf{R}^m$ is such that

$$\forall u\in\mathbf{U}, p\in\mathbf{R}^m : L(u^*,p) \leq L(u^*,p^*) \leq L(u,p^*). \tag{7}$$

From Assumption 1, there exist saddle points of L on $\mathbf{U}\times\mathbf{R}^m$, and we denote the set of saddle points as $\mathbf{U}^*\times\mathbf{P}^*$. By definition, saddle point $(u,p)\in\mathbf{U}^*\times\mathbf{P}^*$ of L satisfies

$$\begin{cases} 0 \in \partial_u L(u,p) + \mathcal{N}_{\mathbf{U}}(u); \\ 0 = -\nabla_p L(u,p). \end{cases} \tag{8}$$

System (8) can also be considered the Karush-Kuhn-Tucker (KKT) system of (P). Thus, the saddle point problem of (P) can be represented as the inclusion problem

$$0 \in H(w) = \begin{pmatrix} \partial_u L(u,p) + \mathcal{N}_{\mathbf{U}}(u) \\ -\nabla_p L(u,p) \end{pmatrix} = \begin{pmatrix} \nabla G(u) + \partial J(u) + A^\top p + \mathcal{N}_{\mathbf{U}}(u) \\ b - Au \end{pmatrix},$$

where, we call H the KKT mapping for obvious reasons.

3. Convergence and convergence rate analysis of RPDC

This section establishes almost surely convergence and expected $O(1/t)$ convergence rate for RPDC. First, we introduce the following assumption on core function K and parameters ϵ and ρ:

Assumption 2 (i) *K is strongly convex with parameter β and differentiable with its gradient Lipschitz continuous with parameter B on $\mathbf{U}$.*

(ii) *Parameters ϵ and ρ satisfy:*

$$0 < \epsilon < \beta/[B_G + \gamma\lambda_{\max}(A^\top A)] \text{and } 0 < \rho < \frac{2\gamma}{2N-1}, \tag{9}$$

where $\lambda_{\max}(A^\top A)$ is the largest eigenvalue of $A^\top A$.

Let $D(u,v) = K(u) - K(v) - \langle \nabla K(v), u - v\rangle$ be a Bregman like function (core function) (Beck & Teboulle, 2003; Cohen & Zhu, 1984). Two popular core functions K satisfy Assumption 2:

1. $K(u) = \frac{1}{2}\|u\|^2$, where $\beta = B = 1$, and
2. $K(u) = \frac{1}{2}\|u\|_Q^2$, where Q is the Q-quadratic norm associated with positive definite matrix Q.

From Assumption 2, $\frac{\beta}{2}\|u-v\|^2 \le D(u,v) \le \frac{B}{2}\|u-v\|^2$ and Algorithm 1 shows the proposed RPDC method to solve (P). For the sake of brevity, let us set $q^k = p^k + \gamma(Au^k - b)$. Then the primal problem can be expressed as

$$(\text{AP}^k)\ \min_{u\in\mathbf{U}} \langle \nabla_{i(k)} G(u^k), u_{i(k)}\rangle + J_{i(k)}(u_{i(k)}) + \langle q^k, A_{i(k)} u_{i(k)}\rangle + \tfrac{1}{\epsilon}[K(u) - \langle \nabla K(u^k), u\rangle].$$

If we choose an additive Bregman function (or core function) with respect to the space decomposition (2), i.e., $K(u) = \sum_{i=1}^{N} K_i(u_i)$, then problem (AP^k) is just a small optimization problem for selected block $i(k)$. Thus, taking $K(u) = \sum_{i=1}^{N} \frac{\|u\|^2}{2}$ for (AP^k), we perform only a block proximal gradient update for block $i(k)$, where we linearize the coupled function $G(u)$ and augmented Lagrangian term $\langle p, Au - b\rangle + \frac{\gamma}{2}\|Au - b\|^2$, and add the proximal term to it.

Algorithm 1 Proposed randomized primal-dual coordinate method

for $k = 1$ **to** t **do**
  Choose $i(k)$ from $\{1, \dots, N\}$ with equal probability;
  $u^{k+1} = \arg\min_{u\in\mathbf{U}} \langle \nabla_{i(k)} G(u^k), u_{i(k)}\rangle + J_{i(k)}(u_{i(k)}) + \langle q^k, A_{i(k)} u_{i(k)}\rangle + \frac{1}{\epsilon} D(u, u^k);$
  $p^{k+1} = p^k + \rho(Au^{k+1} - b).$
end for

Indices $i(k)$, $k = 0, 1, 2, \dots$ in Algorithm 1 are random variables. After k iterations, RPDC generates random output (u^{k+1}, p^{k+1}). We denote $\mathcal{F}_k$ as a filtration generated by the random variable $i(0), i(1), \dots, i(k)$, i.e.,

$$\mathcal{F}_k \stackrel{def}{=} \{i(0), i(1), \dots, i(k)\}, \mathcal{F}_k \subset \mathcal{F}_{k+1},$$

and define $\mathcal{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$, $\mathbb{E}_{\mathcal{F}_{k+1}} = \mathbb{E}(\cdot|\mathcal{F}_k)$ as the conditional expectation with respect to $\mathcal{F}_k$. The conditional expectation in the $i(k)$ term for given $i(0), i(1), \dots, i(k-1)$ is $\mathbb{E}_{i(k)}$.

Let $w = (u,p)$, given $w^* = (u^*, p^*) \in \mathbf{U}^* \times \mathbf{P}^*$. Then for any $w, w' \in \mathbf{U} \times \mathbf{R}^m$, we construct the function

$$\Lambda(w,w') = D(u',u) + \frac{\epsilon}{2N\rho}\|p - p'\|^2 + \frac{\epsilon(N-1)}{N}[L(u,p) - L(u^*,p^*)] + \frac{\epsilon(N-2)\gamma}{2N}\|Au - b\|^2 \tag{10}$$

and we have the following lemma regarding the boundness of $\Lambda(w,w^*)$ and $\Lambda(w,w')$.

Lemma 1 (Boundness of $\Lambda(w,w^*)$ and $\Lambda(w,w')$) *Suppose Assumption 1 and 2 hold. $w^* = (u^*, p^*) \in \mathbf{U}^* \times \mathbf{P}^*$. Then there exist positive numbers d_1, d_2 and d_3, such that*

(i) $\Lambda(w,w^*) \ge d_1\|w - w^*\|^2$,

(ii) $\Lambda(w,w^*) \le d_2\|w - w^*\|^2 + \frac{\epsilon(N-1)}{N}[L(u,p^*) - L(u^*,p^*)]$, *and*

(iii) $\Lambda(w,w') \ge -d_3\|p - p^*\|^2$;

with $d_1 = \min\left\{\frac{1}{2N}[N\beta - \epsilon\gamma\lambda_{\max}(A^\top A)], \frac{\epsilon}{4N\gamma}\right\}$, $d_2 = \max\left\{\frac{(4N-3)\epsilon}{(4N-2)N\rho}, \frac{NB + \epsilon(2N-3)\gamma\lambda_{\max}(A^\top A)}{2N}\right\}$, *and* $d_3 = \frac{\epsilon(N-1)^2}{2\gamma N(N-2)}$.

To analyze the convergence of RPDC, we need the point $T(w^k) = \left(T_u(w^k), T_p(w^k)\right)$ generated by one deterministic iteration of APP-AL for given w^k,

APP-AL

$$\begin{cases} T_u(w^k) = \arg\min_{u\in\mathbf{U}} \langle \nabla G(u^k), u\rangle + J(u) + \langle q^k, Au\rangle \\ \qquad\qquad\qquad\qquad\qquad\qquad + \frac{1}{\epsilon} D(u, u^k); \\ T_p(w^k) = p^k + \gamma\left[AT_u(w^k) - b\right], \end{cases}$$

and the following lemma.

Lemma 2 (Estimation on the variance of $\Lambda(w^k, w)$) *Let Assumption 1 and 2 hold and $\{(u^k, p^k)\}$ be generated by RPDC. Then*

$$\begin{aligned} & \Lambda(w^k, w) - \mathbb{E}_{i(k)}\Lambda(w^{k+1}, w) \\ \geq\ & \frac{\epsilon}{N}\mathbb{E}_{i(k)}\left[L(u^{k+1}, p) - L(u, q^k)\right] + d_4\|w^k - T(w^k)\|^2, \end{aligned}$$

where $d_4 = \frac{\min\left\{\frac{\beta - \epsilon[B_G + \gamma\lambda_{\max}(A^\top A)]}{2}, \frac{\epsilon[2\gamma - (2N-1)\rho]}{2N}\right\}}{\max\{N^2 + 2\gamma^2(N^2+2)\lambda_{\max}(A^\top A), 4\gamma^2\}}$.

3.1. Almost surely convergence of RPDC

Based Lemma 2, we establish the convergence of RPDC.

Theorem 1 (Almost surely convergence) *Let the assumptions of Lemma 2 hold, then*

(i) $\sum_{k=0}^{+\infty} \|w^k - T(w^k)\|^2 < +\infty$ *a.s.*.

(ii) *The sequence $\{w^k\}$ generated by RPDC is almost surely bounded.*

(iii) *Every cluster point of $\{w^k\}$ almost surely is a saddle point of the Lagrangian for (P).*

3.2. Convergence rate analysis for RPDC

This subsection provides the convergence rate of RPDC. We define the average sequence for the sequence $\{(u^k, p^k)\}$ generated from Algorithm RPDC and any $t > 0$ as

$$\bar{u}_t = \frac{\sum_{k=0}^{t} u^{k+1}}{t+1} \text{ and } \bar{p}_t = \frac{\sum_{k=0}^{t} q^k}{t+1}.$$

Theorem 2 (Expected primal suboptimality and expected feasibility)
Let Assumption 1 and 2 hold; (u^*, p^*) be a saddle point, and M be a bound of dual optimal solution of (P), i.e., $\|p^*\| < M$; and $\{(u^k, p^k)\}$ be generated by RPDC. Then we have the following identities.

(i) *Global estimate of expected bifunction values.*

$$\mathbb{E}_{\mathcal{F}_t}\left[L(\bar{u}_t, p) - L(u, \bar{p}_t)\right] \leq \frac{Nh(w^0, w)}{\epsilon(t+1)},$$

$\forall u \in \mathbf{U}, p \in \mathbf{R}^m$, (u,p) could possibly be random and $h(w, w') = \Lambda(w, w') + \frac{d_3}{d_1}\Lambda(w, w^*) \geq 0$.

(ii) *Expected feasibility.*

$$\mathbb{E}_{\mathcal{F}_t}\|A\bar{u}_t - b\| \leq \frac{Nd_5}{(M - \|p^*\|)\epsilon(t+1)},$$

where $d_5 = \sup_{\|p\|<M} h(w^0, (u^*, p))$.

(iii) *Expected suboptimality.*

$$\begin{aligned} \mathbb{E}_{\mathcal{F}_t}\left[F(\bar{u}_t) - F(u^*)\right] &\geq -\frac{\|p^*\|Nd_5}{(M - \|p^*\|)\epsilon(t+1)}, \\ \mathbb{E}_{\mathcal{F}_t}\left[F(\bar{u}_t) - F(u^*)\right] &\leq \frac{Nd_5}{\epsilon(t+1)}. \end{aligned}$$

4. Linear convergence of RPDC under global strong metric subregularity

This section establishes the expected linear convergence of RPDC. Let $\phi(w, w^*) = \Lambda(w, w^*) + \frac{\epsilon}{N}[L(u, p^*) - L(u^*, p^*)]$. Then $\phi(w, w^*) \geq 0$ and $\phi(w, w^*) = 0$ if and only if $w = w^*$, and following lemma holds for the upper bound and descent property of function $\phi(w, w^*)$.

Lemma 3 (Boundness of $\phi(w, w^*)$ and descent inequality) *Suppose Assumption 1 and 2 hold, then the following statements also hold.*

(i) $\phi(w, w^*) \geq d_1\|w - w^*\|^2$.

(ii) $\phi(w, w^*) \leq d_2\|w - w^*\|^2 + \epsilon[L(u, p^*) - L(u^*, p^*)]$.

(iii) $\phi(w^k, w^*) - \mathbb{E}_{i(k)}\phi(w^{k+1}, w^*) \geq d_4\|w^k - T(w^k)\|^2 + \frac{\epsilon}{N}[L(u^k, p^*) - L(u^*, p^*)]$.

The definition for global strong metric subregularity (Dontchev & Rockafellar, 2009) is as follows.

Definition 1 (Global strong metric subregularity) *Let $\mathcal{H}(x)$ be a set-valued mapping between real spaces $\mathbf{X}$ and $\mathbf{Y}$. Then $\mathcal{H}(x)$ is called global strong metric subregular at $\bar{x}$ for $\bar{y}$ when $\bar{y} \in \mathcal{H}(\bar{x})$ if there exists positive number $\mathfrak{c}$ such that*

$$dist(x, \bar{x}) \leq \mathfrak{c}\, dist\left(\bar{y}, \mathcal{H}(x)\right), \quad \textit{for all } x \in \mathbf{X}. \qquad (11)$$

The following theorem regarding linear convergence of RPDC under global strong metric subregularity of the KKT mapping $H : \mathbf{R}^n \times \mathbf{R}^m \rightrightarrows \mathbf{R}^n \times \mathbf{R}^m$.

Theorem 3 (Global strong metric subregularity of $H(w)$ implies linear convergence of RPDC) *Suppose Assumption 1 and 2 hold. For a given saddle point w^*, if $H(w)$ is global strong metric subregular at w^* for 0, then there exists $\alpha \in (0,1)$ such that*

$$\mathbb{E}_{\mathcal{F}_{k+1}}\phi(w^{k+1}, w^*) \leq \alpha^{k+1}\phi(w^0, w^*), \quad \forall k. \tag{12}$$

Then the R-linear of the sequence $\{\mathbb{E}_{\mathcal{F}_k} w^k\}$ can be expressed as in the corollary.

Corollary 1 (R-linear rate of $\{\mathbb{E}_{\mathcal{F}_k} w^k\}$) *Suppose the assumptions of Theorem 3 hold and $\alpha \in (0,1)$ is constant. Then the sequence $\{\mathbb{E}_{\mathcal{F}_k} w^k\}$ converges to the desired saddle point w^* at R-linear rate; i.e.,*

$$\lim_{k\to\infty} \sup \sqrt[k]{\|\mathbb{E}_{\mathcal{F}_k} w^k - w^*\|} = \sqrt{\alpha}.$$

5. Support vector machine and machine learning portfolio problem implementations

This section discusses experiments conducted using MATLAB R2020a on a personal computer with Intel Core i5-6200U CPU (2.40GHz) and 8.00 GB RAM. We also calculate optimal values for all experiments to check the suboptimality of RPDC using the commercial solver CPLEX 12.6.

5.1. Support vector machine problem

Consider the SVM problem,

$$\text{(SVM)} \quad \begin{array}{ll} \min\limits_{u\in[0,c]^n} & \frac{1}{2}u^\top Qu - \mathbf{1}_n^\top u \\ \text{s.t.} & y^\top u = 0 \end{array},$$

where $u \in \mathbb{R}^n$ are the decision variables, and $Q \in \mathbb{R}^{n\times n}$ is a symmetric and positive-definite matrix. Let $Q = (Q_1^\top, Q_2^\top, \cdots, Q_N^\top)^\top \in \mathbf{R}^{n\times n}$ be an appropriate partition of matrix Q and Q_i be an $n_i \times n$ matrix. Then the KKT mapping for SVM is

$$H(w) = \begin{pmatrix} Qu - \mathbf{1}_n + py + \mathcal{N}_{[0,c]^n}(u) \\ y^\top u \end{pmatrix},$$

where $w = (u, p)$. The following proposition shows that the KKT mapping for SVM is global strong metric subregular.

Proposition 1 *Assume there exists at least one component u_i^* of optimal solution u^* that satisfies $0 < u_i^* < c$. Then the KKT mapping for SVM is global strong metric subregular.*

The RPDC scheme with $K(u) = \frac{1}{2}\|u\|^2$ for SVM is

$$\begin{aligned} &\text{Choose } i(k) \text{ from } \{1, 2, \ldots, N\} \text{ with equal probability} \\ &u^{k+1} \leftarrow \min_{u\in[0,c]^n} \langle Q_{i(k)}u^k, u_{i(k)}\rangle - \mathbf{1}_{n_{i(k)}}^\top u_{i(k)} \\ &\qquad + \langle p^k + \gamma y^\top u^k, (y_{i(k)})^\top u_{i(k)}\rangle + \frac{1}{2\epsilon}\|u - u^k\|^2; \\ &p^{k+1} \leftarrow p^k + \rho y^\top u^{k+1}. \end{aligned}$$

Thus, the primal subproblem of RPDC has the closed form

$$\begin{cases} u_{i(k)}^{k+1} = \min\Big\{ \max\Big[0, u_{i(k)}^k - \epsilon\Big(Q_{i(k)}u^k - \mathbf{1}_{n_{i(k)}} \\ \qquad\qquad + (p^k + \gamma y^\top u^k) y_{i(k)}\Big)\Big], c\Big\}, \\ u_{j\neq i(k)}^{k+1} = u_{j\neq i(k)}^k. \end{cases}$$

We used two LIBSVM datasets in the experiment: `heart_scale` (270 data and 13 features) and `ionosphere_scale` (351 data and 34 features). Q was generated using the radial basis function kernel, and we selected $c = 1$.

We partitioned the variables $N = 2, 5, 10$ blocks, for both cases. Thus $n_i = 135, 54, 27$ for the first dataset (`heart_scale`); and $n_i = 175$ (or 176), 70 (or 71), 35 (or 36) for the second dataset (`ionosphere_scale`).

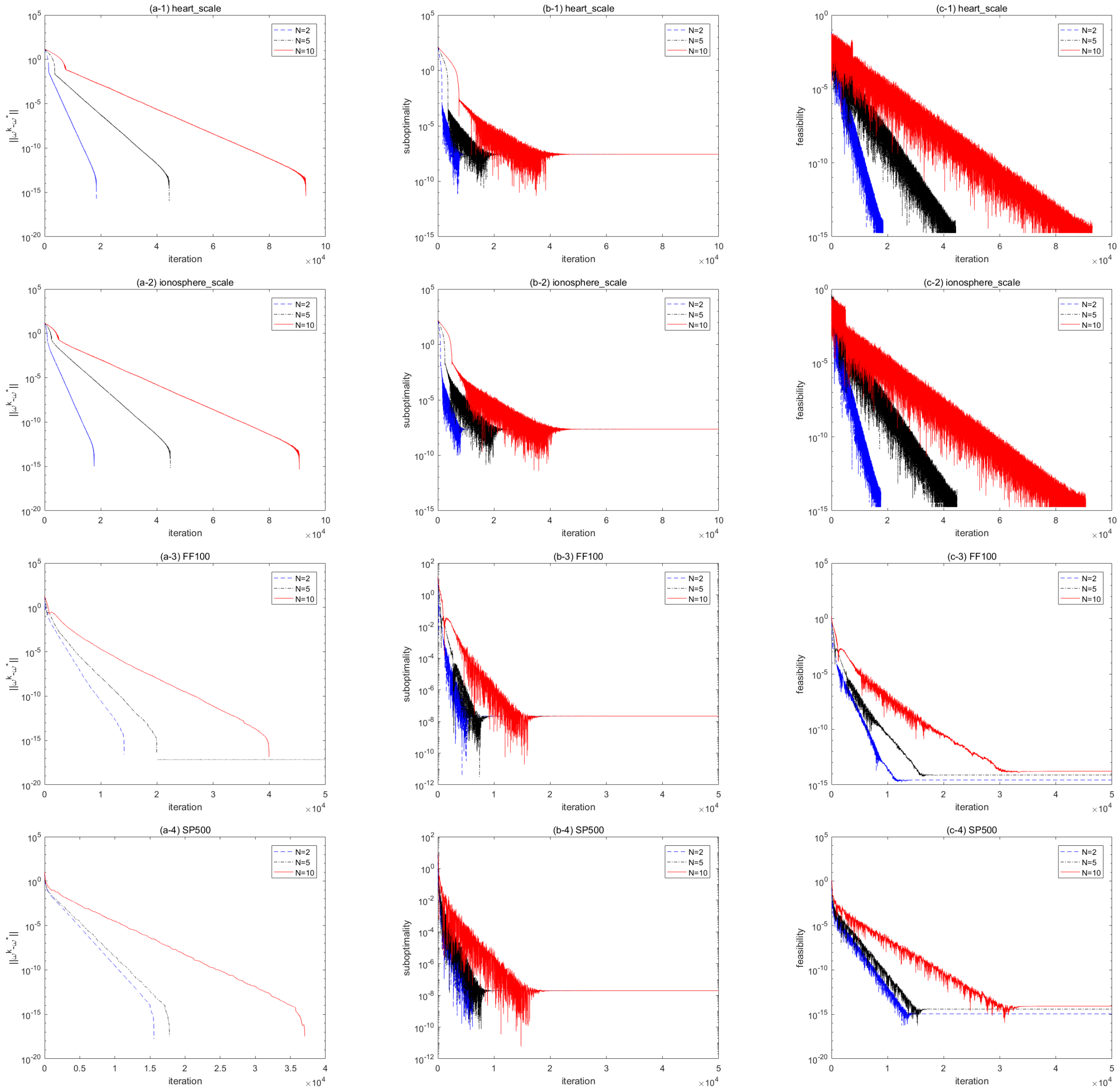
In Figure 1, graphs (a-1) and (a-2) show the number of blocks and $\|w^k - w^*\|$ with respect to iteration count, respectively; graphs (b-1) and (b-2) show the number of blocks and suboptimality with respect to iteration count, respectively; and graphs (c-1) and (c-2) show the number of blocks and feasibility with respect to iteration count, respectively.

We compared three algorithms: APP-AL by (Cohen & Zhu, 1984) and (Zhao & Zhu, 2019), and RPDC from this paper with $N = 2$ and random coordinate descent (RCD) algorithm (Necoara & Patrascu, 2014)) on `heart_scale` and `ionosphere_scale` problems. Suboptimality and feasibility were measured by $|F(u) - F(u^*)| + \|y^\top u\|$ with $F(u) = \frac{1}{2}u^\top Qu - \mathbf{1}_n^\top u$. In Figure 2, graphs (a-1) and (b-1) show $|F(u) - F(u^*)| + \|y^\top u\|$ versus iteration count; and graphs (a-2) and (b-2) show average computation time per iteration for the different algorithms. The total number of iterations required for APP-AL and RPDC are both less than RCD, APP-AL is faster than RPDC. But computation per iteration of RPDC is less than APP-AL.

5.2. Machine learning portfolio problem

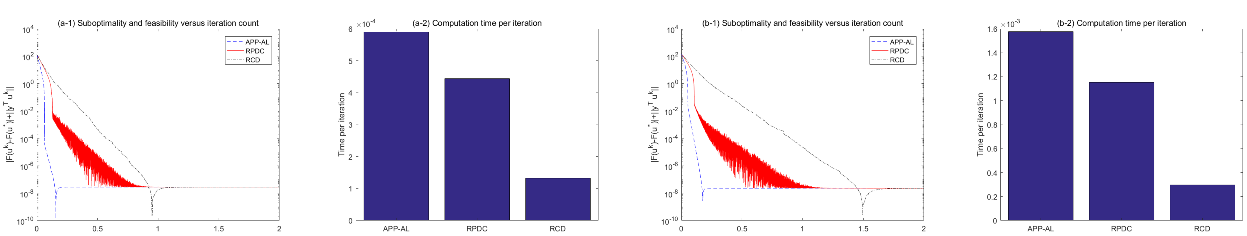
Consider the MLP problem.

$$\text{(MLP)} \quad \begin{array}{ll} \min\limits_{u\in\mathbf{R}^n} & \frac{1}{2}u^\top \Sigma u + \lambda\|u\|_1 \\ \text{s.t.} & \mu^\top u = \rho \\ & \mathbf{1}_n^\top u = 1 \end{array},$$



(a) $\|w^k - w^*\|$ (b) Suboptimality (c) Feasibility

Figure 1. Number of blocks, $\|w^k - w^*\|$, suboptimality, and feasibility with respect to iteration



(a) Comparison for `heart_scale` (b) Comparison for `ionosphere_scale`

Figure 2. Comparing RPDC, APP-AL and RCD

where $u \in \mathbb{R}^n$ is the decision portfolio vector, and $\Sigma \in \mathbb{R}^{n\times n}$, is the symmetric and positive-definite estimated covariance matrix of asset returns. Let $\Sigma = (\Sigma_1^\top, \Sigma_2^\top, \cdots, \Sigma_N^\top)^\top \in \mathbf{R}^{n\times n}$ be an appropriate partition of matrix Σ and Σ_i be an $n_i \times n$ matrix.

The KKT mapping for MLP is:

$$H(w) = \left(\begin{array}{l} \Sigma u + \lambda \partial \|u\|_1 + p_1 \mathbf{1}_n + p_2 \mu \\ \mu^\top u - \rho \\ \mathbf{1}_n^\top u - 1 \end{array} \right),$$

where $w = (u, p)$. The following proposition shows that the KKT mapping for MLP is global strong metric subregular.

Proposition 2 *Assume there exists at least two components u_i^* and u_j^* for optimal solution u^* that satisfy $u_i^* \neq 0$ and $u_j^* \neq 0$; and $\mu_i \neq \mu_j$. Then the KKT mapping for MLP is global strong metric subregular.*

Therefore, the RPDC scheme with $K(u) = \frac{1}{2}\|u\|^2$ for MLP is

$$\begin{aligned} &\text{Choose } i(k) \text{ from } \{1, 2, \ldots, N\} \text{ with equal probability} \\ &u^{k+1} \leftarrow \min_{u \in \mathbf{R}^n} \langle \Sigma_{i(k)} u^k, u_{i(k)} \rangle + \lambda \|u_{i(k)}\|_1 \\ &\qquad + \langle p^k + \gamma \Theta(u^k), \Theta_{i(k)}(u_{i(k)}) \rangle + \frac{1}{2\epsilon}\|u - u^k\|^2; \\ &p^{k+1} \leftarrow p^k + \rho \Theta(u^{k+1}), \end{aligned}$$

where $\Theta(u) = \left(\begin{array}{c} \mu^\top u - \rho \\ \mathbf{1}_n^\top u - 1 \end{array} \right)$ and $\Theta_{i(k)}(u_{i(k)}) = \left(\begin{array}{c} \mu_{i(k)}^\top u_{i(k)} \\ \mathbf{1}_{n_{i(k)}}^\top u_{i(k)} \end{array} \right)$. Thus, the primal subproblem of RPDC has the closed form

$$\left\{ \begin{array}{l} u_{i(k)}^{k+1} = sign(\zeta_{i(k)}^k) \odot \max\{0, |\zeta_{i(k)}^k| - \epsilon\lambda \mathbf{1}_{n_{i(k)}}\}, \\ u_{j\neq i(k)}^{k+1} = u_{j\neq i(k)}^k, \end{array} \right.$$

where
$\zeta_{i(k)}^k = u_{i(k)}^k - \epsilon\left[\Sigma_{i(k)} u^k + \left(\mu_{i(k)}, \mathbf{1}_{n_{i(k)}}\right)(p^k + \gamma\Theta(u^k))\right]$.

Two datasets were chosen to validate RPDC performance.

1. The FF100 dataset from Fama and French benchmark datasets (Fama & French, 1992), created for different financial segments based on data sampled from the U.S. stock market. FF100 formed on the basis of size and book-to-market ratio; and
2. The Standard & Poor's, USA SP500 dataset, November 2004 to April 2016, containing 442 assets and 595 observations.

We partitioned the variables into $N = 2, 5, 10$ blocks for both cases, i.e., $n_i = 50, 20, 10$ and $n_i = 221, 88$ (or 90), 44 (or 46) respectively.

In Figure 1, graphs (a-3) and (a-4) show the number of blocks and $\|w^k - w^*\|$ with respect to iteration count; graphs (b-3) and (b-4) show the number of blocks and suboptimality with respect to iteration; and graphs (c-3) and (c-4) show the number of blocks and feasibility with respect to iteration.

6. Conclusions

This paper proposed a randomized primal-dual coordinate (RPDC) method, a randomized coordinate extension of the first-order primal-dual method proposed by (Cohen & Zhu, 1984) and (Zhao & Zhu, 2019), to solve LCCP. We established almost surely convergence and expected $O(1/t)$ convergence rate for the general convex case, and expected linear convergence under global strong metric subregularity. We showed that SVM and MLP problems satisfy global strong metric subregularity under some reasonable conditions, discussed the implementation details of RPDC, and presented numerical experiments on SVM and MLP problems to verify linear convergence. Future study will consider RPDC for nonlinear convex cone programming with separable and non-separable objective and constraints.

Acknowledgements

This research was supported by NSFC grant 71471112 and 71871140.

References

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152. ACM, 1992.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30): 12267–12272, 2009.

Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

Chang, C.-C. and Lin, C.-J. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

Cohen, G. and Zhu, D. Decomposition and coordination methods in large scale optimization problems: The nondifferentiable case and the use of augmented lagrangians. *Adv. in Large Scale Systems*, 1:203–266, 1984.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Dontchev, A. L. and Rockafellar, R. T. Implicit functions and solution mappings. *Springer Monographs in Mathematics. Springer*, 208, 2009.

Fama, E. F. and French, K. R. The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465, 1992.

Fortin, M. and Glowinski, R. Chapter iii on decomposition-coordination methods using an augmented lagrangian. In *Studies in Mathematics and Its Applications*, volume 15, pp. 97–146. Elsevier, 1983.

Gao, X., Xu, Y.-Y., and Zhang, S.-Z. Randomized primal–dual proximal block coordinate updates. *Journal of the Operations Research Society of China*, 7(2):205–250, 2019.

Hestenes, M. R. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.

Ho, M., Sun, Z., and Xin, J. Weighted elastic net penalized mean-variance portfolio design and computation. *SIAM Journal on Financial Mathematics*, 6(1): 1220–1244, 2015.

Lai, Z.-R., Yang, P.-Y., Fang, L., and Wu, X. Short-term sparse portfolio optimization based on alternating direction method of multipliers. *The Journal of Machine Learning Research*, 19(1):2547–2574, 2018.

Li, B., Sahoo, D., and Hoi, S. C. Olps: a toolbox for on-line portfolio selection. *The Journal of Machine Learning Research*, 17(1):1242–1246, 2016.

Liu, J. and Wright, S. J. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.

Liu, J., Wright, S., Ré, C., Bittorf, V., and Sridhar, S. An asynchronous parallel stochastic coordinate descent algorithm. In *International Conference on Machine Learning*, pp. 469–477, 2014.

Lu, Z. and Xiao, L. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, 2015.

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.

Necoara, I. and Patrascu, A. A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints. *Computational Optimization and Applications*, 57(2):307–337, 2014.

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

Ortega, J. M. and Rheinboldt, W. C. *Iterative solution of nonlinear equations in several variables*, volume 30. Siam, 1970.

Powell, M. J. A method for nonlinear constraints in minimization problems. *Optimization*, pp. 283–298, 1969.

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.

Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000.

Shen, W., Wang, J., and Ma, S. Doubly regularized portfolio with risk minimization. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

Wang, P.-W. and Lin, C.-J. Iteration complexity of feasible descent methods for convex optimization. *The Journal of Machine Learning Research*, 15(1):1523–1548, 2014.

Xu, Y. Asynchronous parallel primal–dual block coordinate update methods for affinely constrained convex programs. *Computational Optimization and Applications*, 72(1):87–113, 2019.

Xu, Y. and Zhang, S. Accelerated primal–dual proximal block coordinate updating methods for constrained convex optimization. *Computational Optimization and Applications*, 70(1):91–128, 2018.

Zhao, L. and Zhu, D. First-order primal-dual method for nonlinear convex cone programming. *arXiv preprint arXiv:1801.00261*, 2019.