DrSVM: Distributed Random Projection Algorithms for SVMs

Soomin Lee
Electrical and Computer Engineering
University of Illinois
Urbana, IL 61801, USA
lee203@illinois.edu

Angelia Nedić
Industrial and Enterprise Systems Engineering
University of Illinois
Urbana, IL 61801, USA
angelia@illinois.edu

***Abstract*— We present distributed random projected gradient algorithms for Support Vector Machines (SVMs) that can be used by multiple agents connected over a time-varying network. The goal is for the agents to cooperatively find the same maximum margin hyperplane. In the primal SVM formulation, the objective function can be represented as a sum of convex functions and the constraint set is an intersection of multiple halfspaces. Each agent minimizes a local objective subject to a local constraint set. It maintains its own estimate sequence and communicates with its neighbors. More specifically, each agent calculates weighted averages of the received estimates and its own estimate, adjust the estimate by using gradient information of its local objective function and project onto a subset of its local constraint set. At each iteration, an agent considers only one halfspace since projection onto a single halfspace is easy. We also consider the convergence behavior of the algorithms and prove that all the estimates of agents converge to the same limit point in the optimal set.**

I. Introduction

Support Vector Machines (SVMs) are popular classification tools with a strong theoretical background. Given a set of example-label pairs $\{(a_j, b_j)\}_{j=1}^n$, $a_j \in \mathbb{R}^d$ and $b_j \in \{+1, -1\}$, we find $x \in \mathbb{R}^d$ that solves the following optimization problem. (A bias term is included in x for convenience.)

$$\begin{aligned} \min \ & f(x) = \frac{1}{2}\|x\|^2 \\ \text{s.t.} \ & b_j\langle x, a_j\rangle \geq 1, \ \forall j \in \{1, \ldots, n\} \end{aligned} \tag{1}$$

If the optimal solution to this problem exists (if the data set is linearly separable), the solution is unique due to the strong-convexity of f and is the maximum-margin hyperplane [1].

There are many SVM algorithms (see `http://www.support-vector-machines.org/` and references therein) but no algorithm solves the problem (1) as it is due to the complicated constraint set. Existing SVM algorithms can be broadly classified into two categories. Algorithms in the first category solve the dual problem of (1) and transform the dual solutions to primal solutions.

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} & W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j}^n \alpha_i \alpha_j b_i b_j a_i^T a_j \\ \text{s.t.} \ & \alpha_i \geq 0, \ \text{for } i = 1, \ldots, n \end{aligned} \tag{2}$$

The second author gratefully acknowledges the NSF support of this work under the grant NSF CMMI 07-42538.

Note that the only constraints here are the nonnegativity constraints. Such a constraint set is easy to handle if we use projection-based algorithms. However, an ϵ-accurate solution of the dual problem does not necessarily imply an ϵ-accurate solution of the primal problem. Also, the formulation (2) is not suitable for applying gradient-based algorithms since $\mathcal{O}(n^2)$ space is required to save the gradient of the function $W(\alpha)$.

Algorithms in the other category solve the problem in the primal space, but they transform the primal problem (1) to the following unconstrained optimization problem with a penalty parameter $\gamma > 0$:

$$\min f(x) = \frac{1}{2}\|x\|^2 + \gamma \sum_{j=1}^n \mathcal{L}(x; (a_j, b_j)) \tag{3}$$

where $\mathcal{L}(x; (a_j, b_j)$ is a non-negative loss function representing the violation of the inequality constraint associated with the data (a_j, b_j). There is a threshold value $\bar{\gamma}$ such that for all $\gamma \geq \bar{\gamma}$, the minimum of (1) coincides with the minimum of (3). However, finding such $\bar{\gamma}$ is not easy.

In this paper, we solve the constrained optimization problem (1) as it is by using gradient descent methods and random feasibility updates proposed in the paper [2]. The random feasibility updates are performed based on projecting onto the random observations of the constraint components. That is, instead of projecting on the whole constraint set, we consider a single constraint observed at a time and project onto that constraint. The algorithm addressed here is therefore, by nature an online algorithm. In the SVMs, each constraint is just a halfspace so projection onto such constraint set is easy (a closed-form solution exists).

Moreover, to deal with scalability issues, we distribute the problem over a network so that multiple agents cooperatively solve the problem with limited information and yet achieve a common goal. We propose to solve the problem in online settings without additional consensus constraints or any approximations of the original problem.

A number of distributed SVM algorithms have been also developed. To name a few, PSVM [3] performs row-based matrix factorization to perform a parallel computation, consensus-based distributed SVM [4] introduces additional consensus constraints to enforce consistency across agents, DSSVM [5] and DPSVM [6] solve a local SVM at each node to figure out support vectors and communicate these among

agents. To the best of our knowledge, none of the existing distributed SVM algorithms are applicable in online settings or solve the problem (1) without additional changes.

This paper is organized as follows. In Section II, we formulate a convex distributed optimization problem, describe our algorithm and discuss how to apply the algorithm on two different SVM formulations. In Section III, we state some results from previous literature that we use in the analysis and study the convergence behavior of our algorithm when applied on the SVM formulations. In Section IV, we present some experimental results on text classification benchmarks. Section V contains some concluding remarks and future directions.

Notation A vector is viewed as a column. We write x^T to denote the transpose of a vector x. The scalar product of two vectors x and y is either $\langle x, y\rangle$ or $x^T y$. We use a superscript i to denote an agent i. For example, x_k^i is the kth iterate of an agent i. We use $\|x\|$ to denote the standard Euclidean norm. We write $\text{dist}(x, \mathcal{X})$ for the distance of a vector x from a closed convex set $\mathcal{X}$, *i.e.*, $\text{dist}(x, \mathcal{X}) = \min_{v\in\mathcal{X}} \|v - x\|$. We use $\Pi_{\mathcal{X}}[x]$ to denote the projection of a vector x on the set $\mathcal{X}$, *i.e.*, $\Pi_{\mathcal{X}}[x] = \arg\min_{v\in\mathcal{X}} \|v-x\|^2$. We use $\mathsf{Pr}\{Z\}$ and $\mathsf{E}[Z]$ to denote the probability and the expectation of a random variable Z. We abbreviate *almost surely* and *independent and identically distributed* as *a.s.* and *iid*, respectively.

II. Distributed Random Projection On SVMs

We consider the problem of optimizing the sum of convex objective functions corresponding to m agents connected over a time-varing topology. The goal of the agents is to cooperatively solve the following optimization problem when the constraint set $\mathcal{X}$ is not known a priori:

$$\min_{x\in\mathbb{R}^d} f(x) = \sum_{i=1}^m f_i(x) \qquad (4)$$
$$\text{s.t. } x \in \mathcal{X} = \bigcap_{i=1}^m \mathcal{X}_i$$

where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is a convex function, representing the local objective function of agent i, and each $\mathcal{X}_i \subseteq \mathbb{R}^d$ is a closed convex set, representing the local constraint set of agent i. We assume the local objective function f_i and the local constraint set $\mathcal{X}_i$ are known to agent i only. In the case of SVM, each $\mathcal{X}_i$ is an intersection of simple sets. *i.e.*,

$$\mathcal{X}_i = \bigcap_{j\in I_i} \mathcal{X}_{ij} \text{ for } i = 1, \ldots, m,$$

where each $\mathcal{X}_{ij}$ is a halfspace. In online settings, an agent i does not know $\mathcal{X}_i$ in advance, but it is revealed to the agent through a random realization of its components $\mathcal{X}_{ij}$ (*i.e.* an inequality associated with one training sample) in time.

To solve the problem, we propose gradient descent methods with random feasibility updates. Each agent i starts with an initial estimate $x_0^i \in \mathcal{X}_i$. At iteration k, an agent i updates its estimate by combining the estimates received from its neighbors, taking a gradient step to minimize its objective function f_i, and projecting on a random realization $\mathcal{X}_{i\Omega_k^i}$ of the sets $\mathcal{X}_{ij}$ to minimize the feasibility violation. Here, $\{\Omega_k^i\}_{k\geq 1}$ is a sequence of iid random variables drawn from the set of indices I_i.

Formally, each agent i updates according to the following rule:

$$v_{k-1}^i = \sum_{j=1}^m w_{k-1}^{ij} x_{k-1}^j \qquad (5a)$$
$$x_k^i = \Pi_{\mathcal{X}_{i\Omega_k^i}}[v_{k-1}^i - \alpha_k \nabla f_i(v_{k-1}^i)], \qquad (5b)$$

where the scalars w_{k-1}^{ij}, for all k, i and j are nonnegative weights and the scalar $\alpha_k > 0$ is a stepsize. We refer to the method (5a)-(5b) as *distributed random projection algorithm* or as DrSVM (Distributed Random Projection on SVM) when applied on SVMs.

A. DrSVM on Linearly Separable Datasets

For a distributed optimization, we define f_i and $\mathcal{X}_i$ in (1) as follows:

$$f_i(x) = \frac{1}{2m}\|x\|^2 \text{ for all } i = 1, \ldots, m, \qquad (6)$$
$$\mathcal{X}_i = \bigcap_{j\in I_i} \mathcal{X}_{ij} = \bigcap_{j\in I_i} \{x \in \mathbb{R}^d \mid b_j\langle x, a_j\rangle \geq 1\}.$$

Here, I_i is defined such that $\bigcup_{i=1}^m I_i = \{1, \ldots, n\}$ and $j \in I_i$ if and only if $\mathcal{X}_i$ contains the inequality constraint associated with the jth training sample (a_j, b_j). Note that each set $\mathcal{X}_{ij}$ is a halfspace, the projection onto which has a closed form solution.

B. DrSVM on Linearly Nonseparable Datasets

If the given data set of example-label pairs $\{(a_j, b_j)\}_{j=1}^n$, $a_j \in \mathbb{R}^d$ and $b_j \in \{+1, -1\}$, is non-separable, a feasible solution to the problem (1) does not exists. Therefore, we change the hard-constraints into soft-constraints by introducing slack variables ξ_j, for $j = 1, \ldots, n$. Let $\boldsymbol{\xi} = [\xi_1 \ \ldots \ \xi_n]^T$. We find $[y^T \ \boldsymbol{\xi}^T] \in \mathbb{R}^{d+n}$ that solves the following optimization problem.

$$\min_{y,\boldsymbol{\xi}} \ f(y, \boldsymbol{\xi}) = \frac{1}{2}\|y\|^2 + C\sum_{j=1}^n \xi_j \qquad (7)$$
$$\text{s.t. } \ b_j\langle y, a_j\rangle \geq 1 - \xi_j, \quad \forall j \in \{1, \ldots, n\}$$
$$\xi_j \geq 0, \quad \forall j \in \{1, \ldots, n\}.$$

Let $x = [y^T \ \boldsymbol{\xi}^T]^T$. For a distributed optimization, we define f_i and $\mathcal{X}_i$ as follows:

$$f_i(y, \boldsymbol{\xi}) = \frac{1}{2m}\|y\|^2 + C\sum_{j\in I_i} \xi_j,$$

$$\mathcal{X}_i = \bigcap_{j\in I_i} \mathcal{X}_{ij} = \bigcap_{j\in I_i} \{x \in \mathbb{R}^{d+n} \mid b_j\langle y, a_j\rangle \geq 1 - \xi_j, \ \xi_j \geq 0\}.$$

Here, I_i is defined such that $\bigcup_{i=1}^m I_i = \{1, \ldots, n\}$; $j \in I_i$ if and only if $\mathcal{X}_i$ contains the inequality constraint associated with the jth training sample (a_j, b_j) and the nonnegativity constraint associated with the slack variable ξ_j. Note that each set $\mathcal{X}_{ij}$ is an intersection of two halfspaces, the projection onto which can be computed in a few steps.

III. Convergence Analysis

A. Assumptions and Definitions

We next discuss assumptions that we use to prove the convergence behavior of the distributed random projection algorithm (5a)-(5b).

The first assumption requires the agents to communicate sufficiently often so that all the component constraint sets directly or indirectly influence the iterate of any agent. We define N_k^i as the set of agents that agent i communicates with at iteration k. Define (V, E_k) to be the graph with nodes $V = \{1, \ldots, m\}$ and edges

$$E_k = \{(j, i) \mid j \in N_k^i,\ i \in V\}.$$

Assumption 1: There exists a scalar Q such that the graph $(V, \bigcup_{\ell=1,\ldots,Q} E_{k+\ell})$ is strongly connected for all k.

Assumption 2: [Weight Rule] There exists a scalar δ with $0 < \eta < 1$ such that for all $i \in \{1, \ldots, m\}$

(a) $w_k^{ii} \geq \delta$ for all $k \geq 0$.
(b) If $w_k^{ij} > 0$, then $w_k^{ij} \geq \delta$.
(c) $w_k^{ij} = 0$ when $j \notin N_k^i$

Assumption 3: [Doubly Stochasticity] The vectors $w_k^i = (w_k^{i1}, \ldots, w_k^{im})^T$ satisfy:

(a) $w_k^i \geq 0$ and $\sum_{j=1}^m w_k^{ij} = 1$ for all i and k.
(b) $\sum_{i=1}^m w_k^{ij} = 1$ for all j and k.

Assumption 3(a) states that each agents takes a convex combination of its own estimate and the estimates of its neighbors. Assumption 3(b) ensures that the estimate of every agent is influenced by the estimates of every other agent.

Let W_k be the matrix with the (i, j)th entry equal to w_k^{ij}. As a consequence of Assumptions 2-3, the matrix W_k is doubly stochastic. Define for all k, s with $k \geq s$,

$$\Phi(k, s) = W_k W_{k-1} \ldots W_{s+1}. \tag{8}$$

$[\Phi(k, s)]_{ij}$ denotes the (i, j)th entry of the matrix $\Phi(k, s)$.

Next, we state the convergence property of $\Phi(k, s)$ (see [7] for details). The convergence is geometric and the rate of convergence is given by

$$\left|[\Phi(k, s)]_{ij} - \frac{1}{m}\right| \leq \theta \beta^{k-s} \qquad \text{for all } i \text{ and } j, \tag{9}$$

and for all $k > s$, where

$$\theta = \left(1 - \frac{\delta}{4m^2}\right)^{-2}, \quad \beta = \left(1 - \frac{\delta}{4m^2}\right)^{\frac{1}{Q}}.$$

For the random sequence $\{\Omega_k^i\}_{k \geq 1}$, we assume the following.

Assumption 4: The sequences $\{\Omega_k^i\}_{k \geq 1}$ are iid and independent of the initial random point x_0^i. We have $\mathsf{Pr}\{\Omega_k^i = j\} > 0$ for all $j \in I_i$ and $i = 1, \ldots, m$.

Next assumption is crucial for our convergence analysis. One can verify it holds by using the results in [8].

Assumption 5: For all $i \in \{1, \ldots, m\}$, for a random variable $\Omega^i \in I_i$, there exists a constant $c > 0$ such that

$$\text{dist}^2(x, \mathcal{X}) \leq c\mathsf{E}\left[\text{dist}^2(x, \mathcal{X}_{i\Omega^i})\right] \text{ for all } x \in \mathbb{R}^d. \tag{10}$$

B. Preliminaries

The objective function f in problem (1) and (7) are differentiable and has Lipschitz gradients with constant L over the set $\mathbb{R}^d$. *i.e.*,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^d.$$

We can easily verify that $L = 1$ for both cases.

Next, we state a projection theorem which will be used in the convergence analysis (see [9] for its proof).

Theorem 6: [Projection Theorem] Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty closed convex set. The function $\Pi_{\mathcal{X}} : \mathbb{R}^d \to \mathcal{X}$ is continuous and nonexpansive, *i.e.*,

$$\|\Pi_{\mathcal{X}}[x] - \Pi_{\mathcal{X}}[y]\| \leq \|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

In the next lemma, we state a result for our convergence analysis of the method (5a)-(5b) (see [2] for its proof). It provides a basic relation between the iterates x_k^i and v_{k-1}^i.

Lemma 7: Let $\mathcal{Y}$ be a closed convex set such that $\mathcal{Y} \subseteq \mathbb{R}^d$. Let y be given by

$$y = \Pi_{\mathcal{Y}}[x - \alpha \nabla f(x)] \quad \text{for some } x \in \mathbb{R}^d \text{ and } \alpha > 0.$$

Also, let the function f be differentiable over the set $\mathbb{R}^d$ with Lipschitz continuous gradients with constant L. Then, we have for any $\bar{x} \in \mathcal{Y}$ and $z \in \mathbb{R}^d$,

$$\begin{aligned}\|y - \bar{x}\|^2 \leq (1 + A_\eta \alpha^2)\|x - \bar{x}\|^2 & \\ - 2\alpha(f(z) - f(\bar{x})) - \frac{3}{4}\|y - x\|^2 & \\ + \left(\frac{3}{8\eta} + 2\alpha L\right)\|x - z\|^2 + B_\eta \alpha^2 \|\nabla f(\bar{x})\|^2, &\end{aligned}$$

where $A_\eta = 8L^2 + 16\eta L^2$, $B_\eta = 8\eta + 8$ and $\eta > 0$ is arbitrary.

We also make use of two lemmas from the paper [7, Lemma 3.1(a), Lemma 4.1]. For a scalar β and a scalar sequence $\{\gamma_k\}$, we consider the convolution sequence $\sum_{\ell=0}^k \beta^{k-\ell}\gamma_\ell$.

Lemma 8: If $\lim_{k\to\infty} \gamma_k = \gamma$ and $0 < \beta < 1$, then

$$\lim_{k\to\infty} \sum_{\ell=0}^k \beta^{k-\ell}\gamma_\ell = \frac{\gamma}{1 - \beta}.$$

The next lemma measures the agent disagreements with respect to the average sequence $\bar{x}_k = \frac{1}{m}\sum_{i=1}^m x_k^i$.

Lemma 9: Let Assumptions 1-3 hold. Also, let the set $\mathcal{X}$ is closed and convex. Define for all $j \in \{1, \ldots, m\}$ and all k,

$$e_{k+1}^j = x_{k+1}^j - v_k^j.$$

Then, for all $i \in V$ and $k \geq 0$,

$$\begin{aligned}\|x_{k+1}^i - \bar{x}_{k+1}\| \leq m\theta\beta^{k+1} \max_j \|x_0^j\| + \|e_{k+1}^i\| & \\ + \sum_{\ell=1}^k \theta\beta^{k-\ell+1} \sum_{j=1}^m \|e_\ell^j\| + \frac{1}{m}\sum_{j=1}^m \|e_{k+1}^j\|. &\end{aligned}$$

For the proof of our distributed random projection algorithm, we use the following supermartingale theorem. (see [10, Lemma 11, p. 50]).

Theorem 10: Let $\{v_k\}$, $\{u_k\}$, $\{a_k\}$ and $\{b_k\}$ be sequences of nonnegative random variables such that

$$\mathsf{E}[v_{k+1}|\mathfrak{F}_k] \leq (1+a_k)v_k - u_k + b_k \text{ for all } k \geq 0 \quad a.s.$$

where $\mathfrak{F}_k$ denotes the collection $v_0, \ldots, v_k$, $u_0, \ldots, u_k$, $a_0, \ldots, a_k$ and $b_0, \ldots, b_k$. Also, let $\sum_{k=0}^{\infty} a_k < \infty$ and $\sum_{k=0}^{\infty} b_k < 0$ a.s. Then, we have $\lim_{k\to\infty} v_k = v$ for a random variable $v \geq 0$ *a.s.*, and $\sum_{k=0}^{\infty} u_k < \infty$ *a.s.*

C. Almost Sure Convergence

We prove our main convergence result for the distributed random projection algorithm on problem (1). As the function f is strongly-convex, an optimal solution always exists and is unique.

Proposition 11: Let Assumptions 1-5 hold. Let x^* be an optimal solution of f. Let $\{x_k^i\}$ be the iterates generated by the algorithm (5a)-(5b). Let the stepsize be such that $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$. Then, the sequences $\{x_k^i\}$, $i = 1, \ldots, m$, almost surely converge to the optimal point x^*. *i.e.*.

$$\lim_{k\to\infty} x_k^i = x^* \quad \text{for all } i \in \{1, \ldots, m\}.$$

Proof: Using Lemma 7 with the following identification: $\mathcal{Y} = \mathcal{X}_{i\Omega_k^i}$, $y = x_k^i$, $x = v_{k-1}^i$, $\bar{\mathcal{Y}} = \mathcal{X}$, $z = z_{k-1}^i := \Pi_{\mathcal{X}}[v_{k-1}^i]$, $L = 1$ and $\eta = c$ where c is the constant from the relation (10). Thus, for all $\bar{x} \in \mathcal{X}$, $k \geq 1$ and $i \in \{1, \ldots, m\}$,

$$\begin{aligned}\|x_k^i - \bar{x}\|^2 \leq{}& (1+A\alpha_k^2)\|v_{k-1}^i - \bar{x}\|^2 \\ &- 2\alpha_k(f_i(z_{k-1}^i) - f_i(\bar{x})) - \frac{3}{4}\|x_k^i - v_{k-1}^i\|^2 \\ &+ \left(\frac{3}{8c} + 2\alpha_k\right)\|v_{k-1}^i - z_{k-1}^i\|^2 + B\alpha_k^2\|\nabla f_i(\bar{x})\|^2,\end{aligned}$$

with $A = 8 + 16c$ and $B = 8c + 8$, where c is the constant from Assumption 5.

Since $v_{k-1}^i = \sum_{j=1}^m w_{k-1}^{ij} x_{k-1}^j$, using the convexity of the norm square function, we have

$$\|v_{k-1}^i - \bar{x}\|^2 = \|\sum_{j=1}^m w_{k-1}^{ij} x_{k-1}^j - \bar{x}\|^2 \leq \sum_{j=1}^m w_{k-1}^{ij}\|x_{k-1}^j - \bar{x}\|^2.$$

Combining the preceding two relations, we obtain for all $\bar{x} \in \mathcal{X}$, $k \geq 1$ and $i \in \{1, \ldots, m\}$,

$$\begin{aligned}\|x_k^i - \bar{x}\|^2 \leq{}& (1+A\alpha_k^2)\sum_{j=1}^m w_{k-1}^{ij}\|x_{k-1}^j - \bar{x}\|^2 \\ &- 2\alpha_k(f_i(z_{k-1}^i) - f_i(\bar{x})) - \frac{3}{4}\|x_k^i - v_{k-1}^i\|^2 \\ &+ \left(\frac{3}{8c} + 2\alpha_k\right)\|v_{k-1}^i - z_{k-1}^i\|^2 + B\alpha_k^2\|\nabla f_i(\bar{x})\|^2.\end{aligned}$$

By summing the preceding relation over $i = 1, \ldots, m$ and using the doubly stochasticity of the weights, *i.e.*,

$$\begin{aligned}\sum_{i=1}^m\sum_{j=1}^m w_{k-1}^{ij}\|x_{k-1}^j - \bar{x}\|^2 &= \sum_{j=1}^m\left(\sum_{i=1}^m w_{k-1}^{ij}\right)\|x_{k-1}^j - \bar{x}\|^2 \\ &= \sum_{j=1}^m \|x_{k-1}^j - \bar{x}\|^2,\end{aligned}$$

we obtain for all $\bar{x} \in \mathcal{X}$ and $k \geq 1$,

$$\begin{aligned}&\sum_{i=1}^m \|x_k^i - \bar{x}\|^2 \leq (1+A\alpha_k^2)\sum_{j=1}^m \|x_{k-1}^j - \bar{x}\|^2 \\ &- 2\alpha_k\sum_{i=1}^m (f_i(z_{k-1}^i) - f_i(\bar{x})) - \frac{3}{4}\sum_{i=1}^m \|x_k^i - v_{k-1}^i\|^2 \\ &+ \left(\frac{3}{8c} + 2\alpha_k\right)\sum_{i=1}^m \|v_{k-1}^i - z_{k-1}^i\|^2 + B\alpha_k^2\sum_{i=1}^m \|\nabla f_i(\bar{x})\|^2.\end{aligned}$$

From the relation of f and f_i, and the convexity of f, we have

$$\begin{aligned}&\sum_{i=1}^m f_i(z_{k-1}^i) = \frac{1}{m}\sum_{i=1}^m f(z_{k-1}^i) \geq f(\frac{1}{m}\sum_{i=1}^m z_{k-1}^i), \\ &\text{and } \sum_{i=1}^m f_i(\bar{x}) = f(\bar{x}).\end{aligned}$$

Also, let $\bar{z}_{k-1} = \frac{1}{m}\sum_{i=1}^m z_{k-1}^i$. Since $z_{i-1}^i \in \mathcal{X}$ for all i, $\bar{z}_{k-1} \in \mathcal{X}$.

Using these relations, we have for all $\bar{x} \in \mathcal{X}$ and $k \geq 1$,

$$\begin{aligned}&\sum_{i=1}^m \|x_k^i - \bar{x}\|^2 \leq (1+A\alpha_k^2)\sum_{j=1}^m \|x_{k-1}^j - \bar{x}\|^2 \\ &- 2\alpha_k(f(\bar{z}_{k-1}) - f(\bar{x})) - \frac{3}{4}\sum_{i=1}^m \|x_k^i - v_{k-1}^i\|^2 \\ &+ \left(\frac{3}{8c} + 2\alpha_k\right)\sum_{i=1}^m \|v_{k-1}^i - z_{k-1}^i\|^2 + B\alpha_k^2\sum_{i=1}^m \|\nabla f_i(\bar{x})\|^2.\end{aligned}$$

By the definition of x_k^i, we have $x_k^i \in \mathcal{X}_{i\Omega_k^i}$, which implies $\|x_k^i - v_{k-1}^i\| \geq \text{dist}(v_{k-1}^i, \mathcal{X}_{i\Omega_k^i})$. Since $z_{k-1}^i = \Pi_{\mathcal{X}}[v_{k-1}^i]$, we have $\|v_{k-1}^i - z_{k-1}^i\| = \text{dist}(v_{k-1}^i, \mathcal{X})$. Using these relations, we obtain for all $\bar{x} \in \mathcal{X}$ and $k \geq 1$,

$$\begin{aligned}&\sum_{i=1}^m \|x_k^i - \bar{x}\|^2 \leq (1+A\alpha_k^2)\sum_{j=1}^m \|x_{k-1}^j - \bar{x}\|^2 \\ &- 2\alpha_k(f(\bar{z}_{k-1}) - f(\bar{x})) - \frac{3}{4}\sum_{i=1}^m \text{dist}^2(v_{k-1}^i, \mathcal{X}_{i\Omega_k^i}) \\ &+ \left(\frac{3}{8c} + 2\alpha_k\right)\sum_{i=1}^m \text{dist}^2(v_{k-1}^i, \mathcal{X}) + B\alpha_k^2\sum_{i=1}^m \|\nabla f_i(\bar{x})\|^2.\end{aligned}$$

By taking the expectation conditioned on $\mathfrak{F}_{k-1}$, and noting that x_{k-1}^i and $\bar{z}_{k-1}$ are fully determined by $\mathfrak{F}_{k-1}$, we have almost surely for all $\bar{x} \in \mathcal{X}$ and $k \geq 1$,

$$\begin{aligned}&\mathsf{E}\left[\sum_{i=1}^m \|x_k^i - \bar{x}\|^2|\mathfrak{F}_{k-1}\right] \leq (1+A\alpha_k^2)\sum_{j=1}^m \|x_{k-1}^j - \bar{x}\|^2 \\ &- 2\alpha_k(f(\bar{z}_{k-1}) - f(\bar{x})) - \frac{3}{4}\sum_{i=1}^m \mathsf{E}\left[\text{dist}^2(v_{k-1}^i, \mathcal{X}_{i\Omega_k^i})|\mathfrak{F}_{k-1}\right] \\ &+ \left(\frac{3}{8c} + 2\alpha_k\right)\sum_{i=1}^m \text{dist}^2(v_{k-1}^i, \mathcal{X}) + B\alpha_k^2\sum_{i=1}^m \|\nabla f_i(\bar{x})\|^2.\end{aligned}$$

Since $v_{k-1}^i \in \mathbb{R}^d$, by relation (10) we have

$$\text{dist}^2(v_{k-1}^i, \mathcal{X}) \leq c\mathsf{E}\left[\text{dist}^2(v_{k-1}^i, \mathcal{X}_{i\Omega_k^i})|\mathfrak{F}_{k-1}\right] \text{ for all } i.$$

Moreover, since $\alpha_k \to 0$, by choosing $\tilde{k}$ large enough so that $2\alpha_k \leq \frac{1}{8c}$, we have for all $k \geq \tilde{k}$,

$$\begin{aligned}
&-\frac{3}{4}\sum_{i=1}^{m}\mathsf{E}\left[\text{dist}^2(v_{k-1}^i, \mathcal{X}_{i\Omega_k^i})|\mathfrak{F}_{k-1}\right] \\
&+\left(\frac{3}{8c}+2\alpha_k\right)\sum_{i=1}^{m}\text{dist}^2(v_{k-1}^i,\mathcal{X}) \\
&\leq -\frac{1}{4c}\sum_{i=1}^{m}\text{dist}^2(v_{k-1}^i,\mathcal{X}).
\end{aligned}$$

Therefore, we have almost surely for all $\bar{x} \in \mathcal{X}$ and $k \geq \tilde{k}$,

$$\begin{aligned}
\mathsf{E}\left[\sum_{i=1}^{m}\|x_k^i-\bar{x}\|^2|\mathfrak{F}_{k-1}\right] &\leq (1+A\alpha_k^2)\sum_{j=1}^{m}\|x_{k-1}^j-\bar{x}\|^2 \\
&- 2\alpha_k(f(\bar{z}_{k-1})-f(\bar{x})) - \frac{1}{4c}\sum_{i=1}^{m}\text{dist}^2(v_{k-1}^i,\mathcal{X}) \\
&+ B\alpha_k^2\sum_{i=1}^{m}\|\nabla f_i(\bar{x})\|^2.
\end{aligned}$$

Now let $\bar{x} = x^*$. Since $z_{k-1}^i = \Pi_{\mathcal{X}}[v_{k-1}^i]$, using $\text{dist}(v_{k-1}^i,\mathcal{X}) = \|v_{k-1}^i - z_{k-1}^i\|$ for all i and the strong convexity of f, we obtain almost surely for all $k \geq \tilde{k}$,

$$\begin{aligned}
\mathsf{E}\left[\sum_{i=1}^{m}\|x_k^i-x^*\|^2|\mathfrak{F}_{k-1}\right] &\leq (1+A\alpha_k^2)\sum_{j=1}^{m}\|x_{k-1}^j-x^*\|^2 \\
&- \alpha_k\|\bar{z}_{k-1}-x^*\|^2 - \frac{1}{4c}\sum_{i=1}^{m}\|v_{k-1}^i-z_{k-1}^i\|^2 \\
&+ B\alpha_k^2\sum_{i=1}^{m}\|\nabla f_i(x^*)\|^2. \qquad (11)
\end{aligned}$$

Since the problem (1) is convex, the gradient mapping $\nabla f(x^*)$ is constant and so is $\nabla f_i(x^*) = \frac{1}{m}\nabla f(x^*)$. Therefore, the relation (11) satisfies all the conditions for the supermartingale theorem. As results of the theorem, we have that the sequence $\{\|x_k^i - x^*\|\}$ is convergent almost surely for $i = 1,\ldots,m$ and

$$\sum_{k=1}^{\infty}\alpha_k\|\bar{z}_{k-1}-x^*\|^2 < \infty \quad a.s.$$

$$\sum_{k=1}^{\infty}\sum_{i=1}^{m}\|v_{k-1}^i-z_{k-1}^i\|^2 < \infty \quad a.s.$$

The condition $\sum_{k=1}^{\infty}\alpha_k = \infty$ and the preceding relations imply that

$$\liminf_{k\to\infty}\|\bar{z}_{k-1}-x^*\|^2 = 0 \quad a.s. \qquad (12a)$$

$$\lim_{k\to\infty}\sum_{i=1}^{m}\|v_{k-1}^i-z_{k-1}^i\|^2 = 0 \quad a.s. \qquad (12b)$$

Next, we prove the following claim:

$$\lim_{k\to\infty}\|x_k^i-\bar{x}_k\| = 0 \ a.s. \ \forall i \in \{1,\ldots,m\}, \qquad (13)$$

where $\bar{x}_k = \frac{1}{m}\sum_{i=1}^{m}x_k^i$.

Define for all $j \in \{1,\ldots,m\}$ and all k,

$$e_{k+1}^j = x_{k+1}^j - v_k^j. \qquad (14)$$

We relate the sequences $\{x_k^i\}$ and $\{v_k^i\}$. Since $\mathcal{X} \subseteq \mathcal{X}_{i\Omega_k^i}$ and $z_k^i \in \mathcal{X}$, we have $z_k^i \in \mathcal{X}_{i\Omega_k^i}$. Using this and the projection theorem 6, the following relation holds for all $i \in \{1,\ldots,m\}$.

$$\begin{aligned}
\|e_{k+1}^i\| = \|x_{k+1}^i - v_k^i\| &= \|\Pi_{\mathcal{X}_{i\Omega_k^i}}[v_k^i - \alpha_k\nabla f_i(v_k^i)] - v_k^i\| \\
&\leq \|\Pi_{\mathcal{X}_{i\Omega_k^i}}[v_k^i - \alpha_k\nabla f_i(v_k^i)] - z_k^i\| + \|z_k^i - v_k^i\| \\
&\leq 2\|z_k^i - v_k^i\| + \alpha_k\|\nabla f_i(v_k^i)\|. \qquad (15)
\end{aligned}$$

From (6), we have $\|\nabla f_i(v_k^i)\| = \frac{1}{m}\|v_k^i\|$. Also, from the relation (5a) and the result that $\{\|x_k^i - x^*\|\}$ is convergent for all $i = 1,\ldots,m$ *a.s.*, we know that both the sequences $\{x_k^i\}$ and $\{v_k^i\}$ are bounded for all $i = 1,\ldots,m$ *a.s.* Therefore, $\|\nabla f_i(v_k^i)\| = \frac{1}{m}\|v_k^i\|$ is bounded *a.s.* As $\alpha_k \to 0$, this and the relation (12b) yield

$$\lim_{k\to\infty}\|e_{k+1}^i\| = 0 \quad a.s. \quad \forall i \in \{1,\ldots,m\} \qquad (16)$$

Next, we use lemma 9. The result is reproduced here for convenience.

$$\begin{aligned}
\|x_{k+1}^i - \bar{x}_{k+1}\| &\leq m\theta\beta^{k+1}\max_j\|x_0^j\| + \|e_{k+1}^i\| \\
&+ \sum_{\ell=1}^{k}\theta\beta^{k-\ell+1}\sum_{j=1}^{m}\|e_\ell^j\| + \frac{1}{m}\sum_{j=1}^{m}\|e_{k+1}^j\|.
\end{aligned}$$

Recall that $\beta = \left(1-\frac{\eta}{4m^2}\right)^{\frac{1}{Q}}$ is a network constant. As $0 < \beta < 1$, the first term on the right hand side converges to zero as $k \to \infty$. From the result (16), we have $\lim_{\ell\to\infty}\sum_{j=1}^{m}\|e_\ell^j\| = 0$. Therefore, from Lemma 8, the third term also converges to zero. Thus, the claim (13) is proved.

From the relations $\bar{z}_k = \frac{1}{m}\sum_{i=1}^{m}z_k^i$, $\bar{v}_k = \frac{1}{m}\sum_{i=1}^{m}v_k^i$, and the relation (12b), we obtain that

$$\lim_{k\to\infty}\|\bar{v}_k - \bar{z}_k\| = 0 \quad a.s. \qquad (17)$$

From the doubly stochasticity of the weights, we have $\bar{x}_k = \bar{v}_k$ for all k. Also, from (12b) and the fact that $\{\|\bar{x}_k - x^*\|\}$ and $\{\|\bar{v}_k - x^*\|\}$ are convergent, $\{\|\bar{z}_k - x^*\|\}$ is convergent. From this, (17), (12a) and (13), the sequences $\{x_k^i\}$, $i = 1,\ldots,m$, almost surely converge to the optimal point x^*, *i.e.*,

$$\lim_{k\to\infty}x_k^i = x^* \quad \text{for all } i \in \{1,\ldots,m\} \quad a.s.$$

■

Due to the lack of space, we do not provide a convergence result for problem (7), which is very similar to the result of problem (1). The only difference is that the sequences $\{x_k^i\}$, $i = 1,\ldots,m$, almost surely converge to an optimal (random) point $x^* \in \mathcal{X}^*$ when problem (7) has a nonempty optimal set $\mathcal{X}^*$.

TABLE I

THE STATISTICS OF THREE TEXT CLASSIFICATION DATA SETS: n IS THE NUMBER OF EXAMPLES AND d IS THE NUMBER OF FEATURES. s REPRESENTS THE SPARSITY OF DATA.

Data set	Statistics		
	n	d	s
astro-ph	62,369	99,757	0.08%
CCAT	804,414	47,236	0.16%
C11	804,414	47,236	0.16%

TABLE II

THE RESULTS OF DRSVM WITH THREE DIFFERENT NUMBER OF AGENTS ($m = 2, 5, 10$): n_i IS THE NUMBER OF ITERATIONS TO REACH THE STOPPING CRITERIA; e_{gen} IS THE GENERALIZATION ERROR OF THE FINAL SOLUTION.

Data set	$m = 2$		$m = 5$		$m = 10$	
	n_i	e_{gen}	n_i	e_{gen}	n_i	e_{gen}
astro-ph	219	0.10	972	0.05	4136	0.04
CCAT	196	0.14	1751	0.07	5118	0.06
C11	107	0.03	931	0.03	3674	0.03

IV. EXPERIMENTAL RESULTS

In the section, we perform some experiments with our distributed random projection algorithm. The purpose of the experiments is to verify the convergence and to show in how many iterations the proposed method can actually arrive at consensus in distributed settings. For simplicity, we use the algorithm (5a)-(5b) with $\alpha_k = \frac{1}{k}$ and $w_k^{ij} = \frac{1}{m}$ for all k, i and j.

We use 3 text classification data sets for our experiments. The data sets were kindly provided by Thorsten Joachims (see [11] for their descriptions). Table I lists the statistics of the data sets. All of the data sets are from binary document classification. Since the data sets used here are very unlikely separable, we use the formulation (7) with $C = 1$. To estimate the generalization (or testing) performance, we split the data and use 80% for training and 20% for testing.

DrSVM is implemented with C/C++ and all experiments were performed on a 64-bit machine running Fedora 16 with an Intel Core 2 Quad Processor Q9400 and 8G of RAM. The experiments are not performed on a real networked environment so we do not consider delays and node/link failures that may exist on networks.

For stopping criteria, we use i) the relative error between an average estimate and an estimate from each agent and ii) the relative error of objective values in two consecutive iterations. *i.e.*,

$$\frac{\|\bar{x}_k - x_k^i\|}{\|\bar{x}_k\|} \leq 0.01, \quad \forall i \in \{1, \ldots, m\},$$

where $\bar{x}_k = \frac{1}{m}\sum_{i=1}^m x_k^i$ and

$$|f(x_k^i) - f(x_{k+1}^i)|/f(x_k^i) < 0.001, \quad \forall i \in \{1, \ldots, m\}.$$

If both criteria are satisfied, we consider that the agents arrived at consensus and the algorithm converged at iteration k.

Table II shows the results. As the number of agents increases, it seems that more iterations are needed for DrSVM to converge. This is because the stopping criteria become harder to be satisfied. In general, we can observe that the algorithm converge very quickly for all three data sets. For example, CCAT with $m = 10$ requires 5118 iterations for convergence. This means DrSVM performs at most 51,180 projections, which is much less than the number of training examples (804,414).

V. CONCLUSIONS AND FUTURE WORK

We studied distributed random projection algorithms on SVMs for a network of agents with time-varying connectivity. Each agent i's estimate is constrained on a closed convex set $\mathcal{X}_i$ and in the case of SVM, the projection onto the subset of $\mathcal{X}_i$ is simple. Under some assumptions on the network connectivity, we proved that each of the estimates converge to the same limit point in $\mathcal{X} = \bigcap_{i=1}^m \mathcal{X}_i$. Experiments on three text classification benchmarks were performed to verify the performance of the proposed algorithms.

There are a few directions to extend this research. First, we can extend this idea to nonlinear SVMs to deal with more general types of data sets. Second, more robust algorithms can be developed to also handle asynchronous networks with communication delays, noise and/or failures in links/nodes. Third, an implementation of a real parallel computing environment will be needed to handle large-scale data sets.

REFERENCES

[1] V. N. Vapnik, *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[2] A. Nedić, "Random algorithms for convex minimization problems," *Mathematical Programming - B*, vol. 129, pp. 225–253, 2011.

[3] E. Y. Chang, K. Zhu, H. Wang, H. Bai, J. Li, Z. Qiu, and H. Cui, "PSVM: Parallelizing Support Vector Machines on Distributed Computers," in *NIPS*, 2007.

[4] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *J. Mach. Learn. Res.*, vol. 99, pp. 1663–1707, August 2010.

[5] A. Navia-Vazquez, D. Gutierrez-Gonzalez, E. Parrado-Hernandez, and J. Navarro-Abellan, "Distributed support vector machines," *IEEE Transactions on Neural Networks*, vol. 17, pp. 1091 – 1097, July 2006.

[6] Y. Lu, V. Roychowdhury, and L. Vandenberghe, "Distributed support vector machines," *IEEE Transactions on Neural Networks*, vol. 19, pp. 1167 – 1178, July 2008.

[7] S. S. Ram, A. Nedich, and V. V. Veeravalli, "Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization," *Journal of Optimization Theory and Applications*, vol. 147, pp. 516–545, 2010.

[8] J. V. Burke and M. C. Ferris, "Weak sharp minima in mathematical programming," *SIAM Journal on Control and Optimization*, vol. 31, pp. 1340–1359, 1993.

[9] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex analysis and optimization*. Athena Scientific, 2003.

[10] B. Polyak, *Intro. to optimization*. Optimization software, Inc., Publications division, New York, 1987.

[11] T. Joachims, "Training linear svms in linear time," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 217–226.