Projection-free Distributed Online Learning in Networks

Wenpeng Zhang [1] **Peilin Zhao** [2] **Wenwu Zhu** [1] **Steven C. H. Hoi** [3] **Tong Zhang** [4]

Abstract

The conditional gradient algorithm has regained a surge of research interest in recent years due to its high efficiency in handling large-scale machine learning problems. However, none of existing studies has explored it in the distributed online learning setting, where locally light computation is assumed. In this paper, we fill this gap by proposing the distributed online conditional gradient algorithm, which eschews the expensive projection operation needed in its counterpart algorithms by exploiting much simpler linear optimization steps. We give a regret bound for the proposed algorithm as a function of the network size and topology, which will be smaller on smaller graphs or "well-connected" graphs. Experiments on two large-scale real-world datasets for a multiclass classification task confirm the computational benefit of the proposed algorithm and also verify the theoretical regret bound.

1. Introduction

The conditional gradient algorithm (Frank & Wolfe, 1956) (also known as Frank-Wolfe) is historically the earliest algorithm for solving general constrained convex optimization problems. Due to its projection-free property and ability to handle structural constraints, it has regained a significant amount of interest in recent years. Different properties concerning the algorithm, such as the sparsity property (Clarkson, 2010) and the primal-dual convergence rate (Jaggi, 2013), have been analyzed in details. Many different algorithm variants, such as the composite variant (Harchaoui et al., 2015), the online and stochastic variants (Hazan & Kale, 2012) (Hazan, 2016) (Hazan & Luo, 2016), faster variants over special types of convex domains, i.e. spectrahedron (Garber, 2016) and polytope (Garber & Hazan, 2016), have also been proposed.

However, despite this great flourish of research on conditional gradient, its distributed online variant over networks has rarely been investigated. In comparison to this situation is the popularity of the variants of its gradient descent and dual averaging counterparts—distributed online gradient descent (D-OGD) (Ram et al., 2010) (Yan et al., 2013) and distributed online dual averaging (D-ODA) (Duchi et al., 2012) (Hosseini et al., 2013) (Lee et al., 2015), which have been successfully applied in handling large-scale streaming data in decentralized computational architectures (*e.g.*, sensor networks and smart phones). Despite the success of these algorithms, the projection operation required in them still limits their further applicability in many settings of practical interests. For example, in matrix learning (Dudík et al., 2012), multiclass classification (Hazan & Luo, 2016) and many other related problems, where the convex domain is the set of all matrices with bounded nuclear norm, the projection operation amounts to computing the full singular value decomposition (SVD) of a matrix, too expensive an operation that does not meet the need of locally light computation in distributed online learning. To avoid this kind of expensive operation, the distributed online variant of conditional gradient is desired since it is expected to be able to eschew the projection operation by using a linear minimization step instead. In the aforementioned case, the linear step amounts to finding the top singular vector of a matrix, which is at least one order of magnitude simpler. However, how to design and analyze this variant remains an open problem.

To fill this gap, in this work, we present the distributed online conditional gradient (D-OCG) algorithm as the desired variant. This algorithm is a novel extension of the previous online variant (Hazan, 2016), in which each local learner communicates its dual variables with its neighbors to cooperate with each other. We present highly non-trivial analysis of the regret bound for the proposed algorithm, which can capture the intuition that the algorithm's regret bound should be larger on graphs with more nodes and should be smaller on "well-connected" graphs (*e.g.*, complete graphs) than on "poorly connected" graphs (*e.g.*, cycle graphs). We

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China [2]Artificial Intelligence Department, Ant Financial Services Group, Hangzhou, China [3]School of Information Systems, Singapore Management University, Singapore [4]Tencent AI Lab, Shenzhen, China. Correspondence to: Wenwu Zhu <wwzhu@tsinghua.edu.cn>, Wenpeng Zhang <zhangwenpeng0@gmail.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

evaluate the performance of the D-OCG algorithm on two large-scale real-world datasets for a multiclass classification task. The experimental results show that D-OCG runs significantly faster than both D-OGD and D-ODA. This illustrates that although the regret bound for D-OCG is in higher order $O(T^{3/4})$ than those in order $O(T^{1/2})$ for D-OGD and D-ODA, its lower computational cost per iteration outweighs the increased number of iterations, making it overall a faster algorithm. The theoretical results regarding the algorithm's regret bound for different graphs are also well confirmed.

2. Preliminaries

In this section, we first give a formal definition of the distributed online convex optimization problem, and then review the two algorithms that our algorithm are built upon.

2.1. Distributed Online Convex Optimization

Let $G = (V, E)$ denote an undirected graph with vertex set $V = \{1, \cdots, n\}$ and edge set $E \subset V \times V$. In distributed online convex optimization defined over G (see the book (Sayed et al., 2014) and the survey (Hazan, 2016) for more details), each node $i \in V$ is associated with a separate agent or learner. At each round $t = 1, \cdots, T$, each learner $i \in V$ is required to generate a decision point $\boldsymbol{x}_i(t)$ from a convex compact set $\mathcal{K} \subseteq \mathbb{R}^m$. Then the adversary replies each learner's decision with a convex loss function $f_{t,i} : \mathcal{K} \to \mathbb{R}$ and each learner suffers the loss $f_{t,i}(\boldsymbol{x}_i(t))$. The communication between learners is specified by the graph G: learner i can only communicate directly with its immediate neighbors $N(i) = \{j \in V | (i, j) \in E\}$. The goal of the learners is to generate a sequence of decision points $\boldsymbol{x}_i(t)$, $i \in V$ so that the regret with respect to each learner i regarding any fixed decision $\boldsymbol{x} \in \mathcal{K}$ in hindsight,

$$R_T(\boldsymbol{x}_i, \boldsymbol{x}) = \sum_{j=1}^{n} \sum_{t=1}^{T} \left(f_{t,j}(\boldsymbol{x}_i(t)) - f_{t,j}(\boldsymbol{x}) \right),$$

is sublinear in T.

We make the following assumptions: (1) each function $f_{t,i}(\boldsymbol{x})$ is *L-Lipschitz* with respect to the L_2 norm $\|\cdot\|$ on $\mathcal{K}$, i.e. $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$, $|f_{t,i}(\boldsymbol{x}) - f_{t,i}(\boldsymbol{y})| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|$. Note that the Lipschitz condition implies that for any $\boldsymbol{x} \in \mathcal{K}$ and any $\nabla f_{t,i}(\boldsymbol{x}) \in \partial f_{t,i}(\boldsymbol{x})$, we have $\|\nabla f_{t,i}(\boldsymbol{x})\| \leq L$. (2) the Euclidean diameter of $\mathcal{K}$ is bounded by D, i.e. $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$, $\|\boldsymbol{x} - \boldsymbol{y}\| \leq D$.

Two definitions are important for deriving useful properties. We say that a function f is β**-smooth** if $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$, we have

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}) \,,\, \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\beta}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

We say that a function f is σ**-strongly convex** if $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$, we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}) \,,\, \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\sigma}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

A σ-strongly convex function f has a very important property: if $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathcal{K}} f(\boldsymbol{x})$, then for any $\boldsymbol{x} \in \mathcal{K}$, we have

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\sigma}{2} \|\boldsymbol{x} - \boldsymbol{x}^*\|^2.$$

2.2. Distributed Online Dual Averaging

In the distributed online dual averaging algorithm (Duchi et al., 2012) (Hosseini et al., 2013), each learner $i \in V$ maintains a sequence of iterates $\boldsymbol{x}_i(t)$ and a sequence of dual variables $\boldsymbol{z}_i(t)$. At each round $t = 1, \cdots, T$, each learner i first computes the new dual variable $\boldsymbol{z}_i(t+1)$ from a weighted combination of its new subgradient $\boldsymbol{g}_i(t)$ and other dual variables $\{\boldsymbol{z}_j(t), j \in N(i)\}$ received from its neighbors, then updates the iterate $\boldsymbol{x}_i(t+1)$ via a proximal projection $\prod_{\mathcal{K}}^{\psi}(\boldsymbol{z}_i(t+1), \alpha_i(t))$, where $\alpha_i(t)$ is a positive number from a non-increasing sequence, $\psi(\boldsymbol{x}) : \mathcal{K} \to \mathbb{R}$ is a proximal function and $\prod_{\mathcal{K}}^{\psi}(\boldsymbol{z}, \alpha) = \arg\min_{\boldsymbol{x} \in \mathcal{K}} \left\{ \langle \boldsymbol{z} \,,\, \boldsymbol{x} \rangle + \frac{1}{\alpha} \psi(\boldsymbol{x}) \right\}$ denotes the projection onto $\mathcal{K}$ operator specified by $\boldsymbol{z}$, α and $\psi(\boldsymbol{x})$ (see (Nesterov, 2009) (Xiao, 2010) for more details). The communication between learners is modeled by a doubly stochastic symmetric matrix P, which satisfies (1) $P_{ij} > 0$ only if $(i, j) \in E\, (i \neq j)$ or $i = j$; (2) $\forall\, i \in V$, $\sum_{j=1}^{n} P_{ij} = \sum_{j \in N(i)} P_{ij} = 1$ and $\forall\, j \in V$, $\sum_{i=1}^{n} P_{ij} = \sum_{i \in N(j)} P_{ij} = 1$.

Algorithm 1 Distributed Online Dual Averaging (**D-ODA**)

Input: convex set $\mathcal{K}$, maximum round number T,
  parameters $\{\alpha_i(t)\}, \forall\, i \in V$
Initialize: $\boldsymbol{x}_i(1) \in \mathcal{K}$, $\boldsymbol{z}_i(1) = \boldsymbol{0}, \forall\, i \in V$
for $t = 1, \cdots, T$ **do**
  The adversary reveals $f_{t,i}, \forall\, i \in V$
  Compute subgradients $\boldsymbol{g}_i(t) \in \partial f_{t,i}(\boldsymbol{x}_i(t)), \forall\, i \in V$
  for *Each Learner* $i \in V$ **do**
    $\boldsymbol{z}_i(t+1) = \sum_{j \in N(i)} p_{ij} \boldsymbol{z}_j(t) + \boldsymbol{g}_i(t)$
    $\boldsymbol{x}_i(t+1) = \prod_{\mathcal{K}}^{\psi}(\boldsymbol{z}_i(t+1), \alpha_i(t))$
  end for
end for

2.3. Online Conditional Gradient

The standard online conditional gradient algorithm (Hazan & Kale, 2012) (Hazan, 2016) eschews the computational expensive projection operation by using a simple linear optimization step instead and is thus much more efficient for many computationally intensive tasks.

Algorithm 2 Online Conditional Gradient (**OCG**)

Input: convex set $\mathcal{K}$, maximum round number T,
parameters η and $\{\sigma_t\}$
Initialize: $\boldsymbol{x}_1 \in \mathcal{K}$
for $t = 1, \cdots, T$ **do**
  The adversary reveals f_t
  Compute a subgradient $\boldsymbol{g}_t \in \partial f_t(\boldsymbol{x}_t)$
  $F_t(\boldsymbol{x}) = \eta \sum_{i=1}^{t-1} \langle \boldsymbol{g}_i , \boldsymbol{x} \rangle + \|\boldsymbol{x} - \boldsymbol{x}_1\|^2$
  $\boldsymbol{v}_t = \arg\min_{\boldsymbol{x} \in \mathcal{K}} \{\langle \nabla F_t(\boldsymbol{x}_t) , \boldsymbol{x} \rangle\}$
  $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \sigma_t(\boldsymbol{v}_t - \boldsymbol{x}_t)$
end for

3. Distributed Online Conditional Gradient

In this section, we first present the proposed distributed online conditional gradient algorithm, and then give the theoretical analysis of its regret bound.

3.1. Algorithm

Algorithm 3 Distributed Online Conditional Gradient **(D-OCG)**

Input: convex set $\mathcal{K}$, maximum round number T,
parameters $\{\eta_i\}$ and$\{\sigma_{t,i}\}$, $\forall\, i \in V$
Initialize: $\boldsymbol{x}_i(1) \in \mathcal{K}$, $\boldsymbol{z}_i(1) = \boldsymbol{0}$, $\forall\, i \in V$
for $t = 1, \cdots, T$ **do**
  The adversary reveals $f_{t,i}, \forall\, i \in V$
  Compute subgradients $\boldsymbol{g}_i(t) \in \partial f_{t,i}(\boldsymbol{x}_i(t)), \forall\, i \in V$
  for *Each Learner* $i \in V$ **do**
    $F_{t,i}(\boldsymbol{x}) = \eta_i \langle \boldsymbol{z}_i(t) , \boldsymbol{x} \rangle + \|\boldsymbol{x} - \boldsymbol{x}_1(1)\|^2$
    $\boldsymbol{v}_i(t) = \arg\min_{\boldsymbol{x} \in \mathcal{K}} \{\langle \nabla F_{t,i}(\boldsymbol{x}_i(t)) , \boldsymbol{x} \rangle\}$
    $\boldsymbol{x}_i(t+1) = \boldsymbol{x}_i(t) + \sigma_{t,i}(\boldsymbol{v}_i(t) - \boldsymbol{x}_i(t))$
    $\boldsymbol{z}_i(t+1) = \sum_{j \in N(i)} p_{ij} \boldsymbol{z}_j(t) + \boldsymbol{g}_i(t)$
  end for
end for

3.2. Analysis

Theorem 1. *The D-OCG algorithm with parameters* $\eta_i = \frac{(1-\sigma_2(P))D}{2(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))LT^{3/4}}$ *and* $\sigma_{t,i} = \frac{1}{\sqrt{t}}$ *for any* $i \in V$ *and any* $t = 1, \cdots, T$ *attains the following regret bound*

$$R_T(\boldsymbol{x}_i, \boldsymbol{x}) \leq 8nLDT^{3/4} + \frac{6\sqrt{n}+1-\sigma_2(P)}{4(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))} LDT^{1/4} + \frac{2(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))}{1-\sigma_2(P)} LDT^{3/4},$$

where $\sigma_2(P)$ *denotes the second largest eigenvalue of matrix* P *and* $1 - \sigma_2(P)$ *denotes the corresponding spectral gap value.*

Remark. (1) The regret bound for D-OCG is in the similar order $O(T^{3/4})$ to that of its centralized variant OCG (Hazan, 2016). (2) Since the connectivity of a graph is captured by its spectral gap value $1-\sigma_2(P)$ (Duchi et al., 2012) (Colin et al., 2016): the better the connectivity of a graph is, the larger the spectral gap value will be, it is easy to verify that this theorem captures the intuition that the D-OCG's regret bound will be larger on larger graphs (the regret bound will be larger when the node size n is larger for all T) and will be smaller on "well-connected" graphs than on "poorly connected" graphs (the regret bound will be smaller when the spectral gap value is larger for certain large T).

To analyze the regret bound for D-OCG, we first establish its connection to D-ODA. To this end, we consider the following points

$$\boldsymbol{x}_i^*(t) = \arg\min_{\boldsymbol{x} \in \mathcal{K}} F_{t,i}(\boldsymbol{x}),$$

where $F_{t,i}(\boldsymbol{x})$ are the functions defined in line 8 in D-OCG. Actually, these points are exactly the iterates of D-ODA with regularization $\psi(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_1(1)\|^2$ applied to the following loss functions

$$\tilde{f}_{t,i}(\boldsymbol{x}) = f_{t,i}(\boldsymbol{x} + (\boldsymbol{x}_i(t) - \boldsymbol{x}_i^*(t))).$$

Note that these loss functions are not the same as the original $f_{t,i}(\boldsymbol{x})$ in D-OCG. The reason is that the subgradients used in the aforementioned D-ODA are $\partial f_{t,i}(\boldsymbol{x}_i(t))$ rather than $\partial f_{t,i}(\boldsymbol{x}_i^*(t))$. Indeed, in this algorithm, the subgradients are evaluated at the points $\boldsymbol{x}_i^*(t)$, thus the corresponding loss functions $\tilde{f}_{t,i}(\boldsymbol{x})$ should satisfy

$$\partial \tilde{f}_{t,i}(\boldsymbol{x}_i^*(t)) = \partial f_{t,i}(\boldsymbol{x}_i(t)).$$

This clearly holds by definition of $\tilde{f}_{t,i}(\boldsymbol{x})$.

Based on the above preparation, we can now present the following lemma.

Lemma 1. *For any fixed* $i \in V$*, the following bound holds for any* $j \in V$ *and any* $\boldsymbol{x} \in \mathcal{K}$

$$\sum_{t=1}^{T} (f_{t,j}(\boldsymbol{x}_i^*(t)) - f_{t,j}(\boldsymbol{x})) \leq 2L \sum_{t=1}^{T} \left\|\boldsymbol{x}_j(t) - \boldsymbol{x}_j^*(t)\right\| + \sum_{t=1}^{T} (\tilde{f}_{t,j}(\boldsymbol{x}_i^*(t)) - \tilde{f}_{t,j}(\boldsymbol{x})).$$

Proof. By definition of $\tilde{f}_{t,j}(\boldsymbol{x})$ and the Lipschitz-ness of $f_{t,j}(\boldsymbol{x})$, for any $\boldsymbol{x} \in \mathcal{K}$, we have

$$\left|\tilde{f}_{t,j}(\boldsymbol{x}) - f_{t,j}(\boldsymbol{x})\right| \leq L \left\|\boldsymbol{x}_j(t) - \boldsymbol{x}_j^*(t)\right\|.$$

Then by plugging in two auxiliary terms in each difference

$f_{t,j}(\boldsymbol{x}_i^*(t)) - f_{t,j}(\boldsymbol{x})$, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \left(f_{t,j}(\boldsymbol{x}_i^*(t)) - f_{t,j}(\boldsymbol{x})\right) \\
&= \sum_{t=1}^{T} \left(f_{t,j}(\boldsymbol{x}_i^*(t)) - \tilde{f}_{t,j}(\boldsymbol{x}_i^*(t)) + \tilde{f}_{t,j}(\boldsymbol{x}) - f_{t,j}(\boldsymbol{x})\right. \\
&\quad \left. + \tilde{f}_{t,j}(\boldsymbol{x}_i^*(t)) - \tilde{f}_{t,j}(\boldsymbol{x})\right) \\
&\leq \sum_{t=1}^{T} \left|f_{t,j}(\boldsymbol{x}_i^*(t)) - \tilde{f}_{t,j}(\boldsymbol{x}_i^*(t))\right| \\
&\quad + \sum_{t=1}^{T} \left|\tilde{f}_{t,j}(\boldsymbol{x}) - f_{t,j}(\boldsymbol{x})\right| \\
&\quad + \sum_{t=1}^{T} (\tilde{f}_{t,j}(\mathbf{x}_i^*(t)) - \tilde{f}_{t,j}(\boldsymbol{x})) \\
&\leq 2L \sum_{t=1}^{T} \left\|\boldsymbol{x}_j(t) - \boldsymbol{x}_j^*(t)\right\| + \sum_{t=1}^{T} (\tilde{f}_{t,j}(\boldsymbol{x}_i^*(t)) - \tilde{f}_{t,j}(\boldsymbol{x})).
\end{aligned}$$

□

Notice that the last summation in the above bound is exactly the part of the regret of D-ODA incurred by learner j with respect to learner i. Thus, to proceed, we require lemmas that allow us to relate the iterates $\boldsymbol{x}_i(t)$ to the iterates $\boldsymbol{x}_i^*(t)$. Let $h_{t,i}(\boldsymbol{x}) = F_{t,i}(\boldsymbol{x}) - F_{t,i}(\boldsymbol{x}_i^*(t))$ and $h_{t,i} = h_{t,i}(\boldsymbol{x}_i(t))$. In the following, we first present a lemma that establishes the recursion between $h_{t+1,i}$ and $h_{t,i}$.

Lemma 2. *The following recursion between $h_{t+1,i}$ and $h_{t,i}$ holds for any $i \in V$ and any $t = 1, \cdots, T$*

$$\begin{aligned}
h_{t+1,i} \leq (1 - \sigma_{t,i}) h_{t,i} + \sigma_{t,i}^2 D^2 \\
+ \eta_i \left\|\boldsymbol{z}_i(t+1) - \boldsymbol{z}_i(t)\right\| \sqrt{h_{t+1,i}},
\end{aligned}$$

where $\boldsymbol{z}_i(t)$ denotes the dual variable defined in D-OCG.

Proof. Using the definitions of $h_{t,i}(\boldsymbol{x})$ and $\boldsymbol{x}_i(t+1)$, the fact that $F_{t,i}(\boldsymbol{x})$ are 2-smooth and the boundedness of $\mathcal{K}$, we obtain

$$\begin{aligned}
h_{t,i}(\boldsymbol{x}_i(t+1)) &= F_{t,i}(\boldsymbol{x}_i(t) + \sigma_{t,i}(\boldsymbol{v}_i(t) - \boldsymbol{x}_i(t))) \\
&\quad - F_{t,i}(\boldsymbol{x}_i^*(t)) \\
&\leq F_{t,i}(\boldsymbol{x}_i(t)) - F_{t,i}(\boldsymbol{x}_i^*(t)) \\
&\quad + \sigma_{t,i} \left\langle \nabla F_{t,i}(\boldsymbol{x}_i(t)) \, , \, \boldsymbol{v}_i(t) - \boldsymbol{x}_i(t) \right\rangle \\
&\quad + \sigma_{t,i}^2 \left\|\boldsymbol{v}_i(t) - \boldsymbol{x}_i(t)\right\|^2 \\
&\leq F_{t,i}(\boldsymbol{x}_i(t)) - F_{t,i}(\boldsymbol{x}_i^*(t)) \\
&\quad + \sigma_{t,i} \left\langle \nabla F_{t,i}(\boldsymbol{x}_i(t)) \, , \, \boldsymbol{v}_i(t) - \boldsymbol{x}_i(t) \right\rangle \\
&\quad + \sigma_{t,i}^2 D^2.
\end{aligned}$$

By the optimality of $\boldsymbol{v}_i(t)$, we have

$$\left\langle \nabla F_{t,i}(\boldsymbol{x}_i(t)) \, , \, \boldsymbol{v}_i(t) \right\rangle \leq \left\langle \nabla F_{t,i}(\boldsymbol{x}_i(t)) \, , \, \boldsymbol{x}_i^*(t) \right\rangle.$$

By the convexity of $F_{t,i}(\boldsymbol{x})$, we have

$$\left\langle \nabla F_{t,i}(\boldsymbol{x}_i(t)) \, , \, \boldsymbol{x}_i^*(t) - \boldsymbol{x}_i(t) \right\rangle \leq F_{t,i}(\boldsymbol{x}_i^*(t)) - F_{t,i}(\boldsymbol{x}_i(t)).$$

Putting the above three inequalities together, we obtain

$$\begin{aligned}
h_{t,i}(\boldsymbol{x}_i(t+1)) &\leq F_{t,i}(\boldsymbol{x}_i(t)) - F_{t,i}(\boldsymbol{x}_i^*(t)) \\
&\quad + \sigma_{t,i}(F_{t,i}(\boldsymbol{x}_i^*(t)) - F_{t,i}(\boldsymbol{x}_i(t))) \\
&\quad + \sigma_{t,i}^2 D^2 \\
&= (1 - \sigma_{t,i})(F_{t,i}(\boldsymbol{x}_i(t)) - F_{t,i}(\boldsymbol{x}_i^*(t))) \\
&\quad + \sigma_{t,i}^2 D^2 \\
&= (1 - \sigma_{t,i}) h_{t,i} + \sigma_{t,i}^2 D^2.
\end{aligned}$$

Next, by definition of $h_{t+1,i}$ and the optimality of $\boldsymbol{x}_i^*(t)$, we have

$$\begin{aligned}
h_{t+1,i} &= F_{t+1,i}(\boldsymbol{x}_i(t+1)) - F_{t+1,i}(\boldsymbol{x}_i^*(t+1)) \\
&= F_{t,i}(\boldsymbol{x}_i(t+1)) - F_{t,i}(\boldsymbol{x}_i^*(t+1)) \\
&\quad + (F_{t+1,i}(\boldsymbol{x}_i(t+1)) - F_{t,i}(\boldsymbol{x}_i(t+1))) \\
&\quad - (F_{t+1,i}(\boldsymbol{x}_i^*(t+1)) - F_{t,i}(\boldsymbol{x}_i^*(t+1))) \\
&\leq F_{t,i}(\boldsymbol{x}_i(t+1)) - F_{t,i}(\boldsymbol{x}_i^*(t)) \\
&\quad + (F_{t+1,i}(\boldsymbol{x}_i(t+1)) - F_{t,i}(\boldsymbol{x}_i(t+1))) \\
&\quad - (F_{t+1,i}(\boldsymbol{x}_i^*(t+1)) - F_{t,i}(\boldsymbol{x}_i^*(t+1))).
\end{aligned}$$

Then, by definition of $F_{t+1,i}(\boldsymbol{x})$ and $F_{t,i}(\boldsymbol{x})$, we have

$$F_{t+1,i}(\boldsymbol{x}) - F_{t,i}(\boldsymbol{x}) = \eta_i \left\langle \boldsymbol{z}_i(t+1) - \boldsymbol{z}_i(t), \boldsymbol{x} \right\rangle.$$

Thus,

$$\begin{aligned}
h_{t+1,i} &\leq h_{t,i}(\boldsymbol{x}_i(t+1)) \\
&\quad + \eta_i \left\langle \boldsymbol{z}_i(t+1) - \boldsymbol{z}_i(t), \boldsymbol{x}_i(t+1) \right\rangle \\
&\quad - \eta_i \left\langle \boldsymbol{z}_i(t+1) - \boldsymbol{z}_i(t), \boldsymbol{x}_i^*(t+1) \right\rangle \\
&= h_{t,i}(\boldsymbol{x}_i(t+1)) \\
&\quad + \eta_i \left\langle \boldsymbol{z}_i(t+1) - \boldsymbol{z}_i(t), \boldsymbol{x}_i(t+1) - \boldsymbol{x}_i^*(t+1) \right\rangle \\
&\leq h_{t,i}(\boldsymbol{x}_i(t+1)) \\
&\quad + \eta_i \left\|\boldsymbol{z}_i(t+1) - \boldsymbol{z}_i(t)\right\| \left\|\boldsymbol{x}_i(t+1) - \boldsymbol{x}_i^*(t+1)\right\|.
\end{aligned}$$

The last inequality follows from the Cauchy-Schwarz inequality.

Now, we derive the bound for $\|\boldsymbol{x}_i(t+1) - \boldsymbol{x}_i^*(t+1)\|$. By definition, $F_{t,i}(\boldsymbol{x})$ are 2-strongly convex and $\boldsymbol{x}_i^*(t) = \arg\min_{\boldsymbol{x} \in \mathcal{K}} F_{t,i}(\boldsymbol{x})$. Thus, using the property of strongly convex functions, for any $\boldsymbol{x} \in \mathcal{K}$, we have

$$\left\|\boldsymbol{x} - \boldsymbol{x}_i^*(t)\right\|^2 \leq F_{t,i}(\boldsymbol{x}) - F_{t,i}(\boldsymbol{x}_i^*(t)).$$

Analogously, it is easy to deduce that

$$\|\boldsymbol{x}_i(t+1) - \boldsymbol{x}_i^*(t+1)\| \leq \sqrt{h_{t+1,i}}.$$

Combining this bound and the above two bounds for $h_{t+1,i}$ and $h_{t,i}(\boldsymbol{x}_i(t+1))$ yields the stated recursion. □

To make the above recursion more concrete, it remains to bound the deviation term $\|\boldsymbol{z}_i(t+1) - \boldsymbol{z}_i(t)\|$, which measures the stability of local dual variables over each node.

Lemma 3. *For any $i \in V$ and any $t = 1, \cdots, T$, the dual variables $\boldsymbol{z}_i(t)$ and $\boldsymbol{z}_i(t+1)$ specified in D-OCG satisfy the following bound*

$$\|\boldsymbol{z}_i(t+1) - \boldsymbol{z}_i(t)\| \leq \frac{1+\sigma_2(P)}{1-\sigma_2(P)}\sqrt{n}L + L.$$

Proof. Let P^r denote the r-th power of matrix P and P_{ij}^r denote the j-th entry of the i-th row of P^r. Then, via a bit of algebra, we can get the following generalized recursion

$$\boldsymbol{z}_i(t+1) = \sum_{j=1}^n P_{ij}^{t+1-s}\boldsymbol{z}_j(s) + \sum_{r=s}^{t-1}\sum_{j=1}^n P_{ij}^{t-r}\boldsymbol{g}_j(r) + \boldsymbol{g}_i(t).$$

Clearly, this recursion reduces to the standard dual variable update in D-OCG when $s = t$. Next, since $\boldsymbol{z}_j(1) = 0$, by setting $s = 1$, we can obtain

$$\boldsymbol{z}_i(t+1) = \sum_{r=1}^{t-1}\sum_{j=1}^n P_{ij}^{t-r}\boldsymbol{g}_j(r) + \boldsymbol{g}_i(t).$$

Then by assuming P^0 to be the identity matrix I_n, we have

$$\boldsymbol{z}_i(t+1) - \boldsymbol{z}_i(t) = \sum_{r=1}^{t-1}\sum_{j=1}^n (P_{ij}^{t-r} - P_{ij}^{t-r-1})\boldsymbol{g}_j(r) + \boldsymbol{g}_i(t).$$

Using the fact that $\|\boldsymbol{g}_i(t)\| \leq L$, the properties of norm functions and the symmetry of matrix P, we obtain

$$\begin{aligned}
&\|\boldsymbol{z}_i(t+1) - \boldsymbol{z}_i(t)\| \\
&= \left\|\sum_{r=1}^{t-1}\sum_{j=1}^n (P_{ij}^{t-r} - P_{ij}^{t-r-1})\boldsymbol{g}_j(r) + \boldsymbol{g}_i(t)\right\| \\
&\leq \sum_{r=1}^{t-1}\sum_{j=1}^n \left|P_{ij}^{t-r} - P_{ij}^{t-r-1}\right| \left\|\boldsymbol{g}_j(r)\right\| + \|\boldsymbol{g}_i(t)\| \\
&\leq L\sum_{r=1}^{t-1}\left\|P_i^{t-r} - P_i^{t-r-1}\right\|_1 + L,
\end{aligned}$$

where P_i^r denotes the i-th column of matrix P^r.

Now we try to bound the L_1 norm sum in the above inequality. By plugging in an all-ones column vector and then using the properties of norm functions, we obtain

$$\begin{aligned}
&\sum_{r=1}^{t-1}\left\|P_i^{t-r} - P_i^{t-r-1}\right\|_1 \\
&= \sum_{r=1}^{t-1}\left\|(P_i^{t-r} - \mathbf{1}/n) - (P_i^{t-r-1} - \mathbf{1}/n)\right\|_1 \\
&\leq \sum_{r=1}^{t-1}(\left\|P_i^{t-r} - \mathbf{1}/n\right\|_1 + \left\|P_i^{t-r-1} - \mathbf{1}/n\right\|_1).
\end{aligned}$$

To proceed, we introduce a useful property of stochastic matrices (Duchi et al., 2012). Let $\Delta_n = \{\boldsymbol{x} \in \mathbb{R}^n \,|\boldsymbol{x} \succeq 0, \sum_{i=1}^n x_i = 1\}$ denote the n-dimensional probability simplex. Then for any positive integer $s = 1, \cdots$ and any $\boldsymbol{x} \in \Delta_n$, the following inequality holds

$$\|P^s\boldsymbol{x} - \mathbf{1}/n\|_1 \leq \sigma_2(P)^s\sqrt{n}.$$

Taking $\boldsymbol{x}$ to be the i-th canonical basis vector $\boldsymbol{e}_i$ in $\mathbb{R}^n$, we have

$$\|P^s\boldsymbol{e}_i - \mathbf{1}/n\|_1 \leq \sigma_2(P)^s\sqrt{n}.$$

Note that this inequality also holds for $s = 0$ since

$$\left\|P^0\boldsymbol{e}_i - \mathbf{1}/n\right\|_1 = \frac{2(n-1)}{n} \leq \sqrt{n},$$

for any $n = 1, \cdots$. In addition, it is easy to verify that

$$\|P_i^s - \mathbf{1}/n\|_1 = \|P^s e_i - \mathbf{1}/n\|_1.$$

Thus, we have

$$\begin{aligned}
&\sum_{r=1}^{t-1}(\left\|P_i^{t-r} - \mathbf{1}/n\right\|_1 + \left\|P_i^{t-r-1} - \mathbf{1}/n\right\|_1) \\
&\leq \sum_{r=1}^{t-1}(\sigma_2(P)^{t-r} + \sigma_2(P)^{t-r-1})\sqrt{n} \\
&= \frac{1+\sigma_2(P)}{1-\sigma_2(P)}(1-\sigma_2(P)^{t-1})\sqrt{n} \\
&\leq \frac{1+\sigma_2(P)}{1-\sigma_2(P)}\sqrt{n}.
\end{aligned}$$

The above equation and the last inequality follow respectively from the summation formula of geometric series and the fact that $\sigma_2(P) < 1$ when P is a doubly stochastic matrix (Berman & Plemmons, 1979).

Combining the above together yields the stated bound. □

Combining the results in Lemma 2 and Lemma 3, we can obtain a more concrete recursion between $h_{t+1,i}$ and $h_{t,i}$, and then deduce the bound for $h_{t,i}$.

Lemma 4. *Assume that the parameters η_i and $\sigma_{t,i}$ in D-OCG are chosen such that $\eta_i(\frac{1+\sigma_2(P)}{1-\sigma_2(P)}\sqrt{n}+1)L\sqrt{h_{t+1,i}} \leq \sigma_{t,i}^2 D^2$. Then the following bound for $h_{t,i}$ holds for any $i \in V$ and any $t = 1, \cdots, T$*

$$h_{t,i} \leq 4D^2\sigma_{t,i}.$$

This lemma can be easily proved using mathematical induction and we place its detailed proof in the Appendix. Now we can deduce the bound for the deviation between $\boldsymbol{x}_i(t)$ and $\boldsymbol{x}_i^*(t)$.

Lemma 5. *For any fixed $i \in V$, the iterates $\boldsymbol{x}_i(t)$ and $\boldsymbol{x}_i^*(t)$ satisfy the following bound*

$$\sum_{t=1}^{T} \|\boldsymbol{x}_i(t) - \boldsymbol{x}_i^*(t)\| \leq \frac{8}{3}DT^{3/4}.$$

Proof. As is given in the proof of Lemma 2, for any $\boldsymbol{x} \in \mathcal{K}$, we have

$$\|\boldsymbol{x} - \boldsymbol{x}_i^*(t)\|^2 \leq F_{t,i}(\boldsymbol{x}) - F_{t,i}(\boldsymbol{x}_i^*(t)).$$

It then follows that

$$\begin{aligned}\|\boldsymbol{x}_i(t) - \boldsymbol{x}_i^*(t)\| &\leq \sqrt{F_{t,i}(\boldsymbol{x}_i(t)) - F_{t,i}(\boldsymbol{x}_i^*(t))} \\ &= \sqrt{h_{t,i}} \\ &\leq 2D\sqrt{\sigma_{t,i}} \\ &= 2Dt^{-1/4}.\end{aligned}$$

The last inequality follows from the bound in Lemma 4 and the last equation follows from the definition of $\sigma_{t,i}$. Thus, summing over $t = 1, \cdots, T$, we obtain

$$\begin{aligned}\sum_{t=1}^{T} \|\boldsymbol{x}_i(t) - \boldsymbol{x}_i^*(t)\| &\leq 2D\sum_{t=1}^{T} t^{-1/4} \\ &\leq 2D(1 + \int_1^T t^{-1/4}dt) \\ &= 2D(1 + \frac{4}{3}t^{3/4}\Big|_1^T) \\ &\leq 2D(\frac{4}{3}T^{3/4} - \frac{1}{3}) \\ &= \frac{8}{3}DT^{3/4}.\end{aligned}$$

□

Before proceeding with the final proof of Theorem 1, we present the regret bound of the D-ODA algorithm applied to the loss functions $\tilde{f}_{t,j}(\boldsymbol{x})$. To this end, we first introduce an auxiliary sequence which are composed of the centralized averages of dual variables over all nodes at each iteration

$$\bar{\boldsymbol{z}}(t) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i(t).$$

Note that the dual variables used in the above definition are exactly those specified in D-OCG. Then, as for the deviation between the local dual variable $\boldsymbol{z}_i(t)$ and the global dual variable $\bar{\boldsymbol{z}}(t)$, we have the following lemma.

Lemma 6. *For any $i \in V$ and any $t = 1, \cdots, T$, the dual variable $\mathbf{z}_i(t)$ defined in D-OCG and their averages $\bar{\boldsymbol{z}}(t)$ over all nodes satisfy the following bound*

$$\|\boldsymbol{z}_i(t) - \bar{\boldsymbol{z}}(t)\| \leq \frac{\sqrt{n}L}{1-\sigma_2(P)}.$$

Two similar bounds for $\|\boldsymbol{z}_i(t) - \bar{\boldsymbol{z}}(t)\|$ in D-ODA are reported in (Duchi et al., 2012) (Hosseini et al., 2013)[1]. Our bound is tighter than both of them. The proof is a little bit similar to that in Lemma 3 and is presented in detail in the Appendix.

We can now give the regret bound for the D-ODA algorithm in the following lemma.

Lemma 7. *The D-ODA algorithm with regularization $\psi(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_1(1)\|^2$ and parameters $\alpha_i(t) = \eta$, $\forall\, i \in V$, applied to the loss functions $\tilde{f}_{t,j}(\boldsymbol{x})$ attains the following regret bound*

$$R_T^a(\boldsymbol{x}_i, \boldsymbol{x}) \leq \frac{6\sqrt{n}+1-\sigma_2(P)}{2(1-\sigma_2(P))}\eta L^2 T + \frac{1}{\eta}D^2.$$

Using our tighter bound for $\|\boldsymbol{z}_i(t) - \bar{\boldsymbol{z}}(t)\|$, this lemma can be easily deduced from the general regret bound for the D-ODA algorithm (Hosseini et al., 2013). The detailed proof is presented in the Appendix.

Now, we are ready to prove our theorem.

Proof of Theorem 1. By plugging in two auxiliary terms in each difference $f_{t,j}(\boldsymbol{x}_i(t)) - f_{t,j}(\boldsymbol{x})$, we have

$$\begin{aligned}&\sum_{t=1}^{T} (f_{t,j}(\boldsymbol{x}_i(t)) - f_{t,j}(\boldsymbol{x})) \\ &= \sum_{t=1}^{T} (f_{t,j}(\boldsymbol{x}_i(t)) - f_{t,j}(\boldsymbol{x}_i^*(t)) + f_{t,j}(\boldsymbol{x}_i^*(t)) - f_{t,j}(\boldsymbol{x})) \\ &\leq \sum_{t=1}^{T} |f_{t,j}(\boldsymbol{x}_i(t)) - f_{t,j}(\boldsymbol{x}_i^*(t))| \\ &\quad + \sum_{t=1}^{T} (f_{t,j}(\boldsymbol{x}_i^*(t)) - f_{t,j}(\boldsymbol{x})).\end{aligned}$$

Using the Lipschitz-ness of $f_{t,j}(\boldsymbol{x})$, we can obtain the following bound for the first summation

$$\sum_{t=1}^{T} |f_{t,j}(\boldsymbol{x}_i(t)) - f_{t,j}(\boldsymbol{x}_i^*(t))| \leq L\sum_{t=1}^{T} \|\boldsymbol{x}_i(t) - \boldsymbol{x}_i^*(t)\|.$$

[1] Strictly, the norm utilized in them is the general dual norm.

Recall that Lemma 1 provides the bound for the second summation. Combining these two bounds together, we have

$$\sum_{t=1}^{T}(f_{t,j}(\boldsymbol{x}_i(t))-f_{t,j}(\boldsymbol{x}))$$
$$\leq L\sum_{t=1}^{T}\|\boldsymbol{x}_i(t)-\boldsymbol{x}_i^*(t)\|+2L\sum_{t=1}^{T}\left\|\boldsymbol{x}_j(t)-\boldsymbol{x}_j^*(t)\right\|$$
$$+\sum_{t=1}^{T}(\tilde{f}_{t,j}(\boldsymbol{x}_i^*(t))-\tilde{f}_{t,j}(\boldsymbol{x})).$$

Hence,

$$\begin{aligned}R_T(\boldsymbol{x}_i,\boldsymbol{x}) &= \sum_{j=1}^{n}\sum_{t=1}^{T}(f_{t,j}(\boldsymbol{x}_i(t))-f_{t,j}(\boldsymbol{x}))\\ &\leq nL\sum_{t=1}^{T}\|\boldsymbol{x}_i(t)-\boldsymbol{x}_i^*(t)\|\\ &\quad+2L\sum_{j=1}^{n}\sum_{t=1}^{T}\left\|\boldsymbol{x}_j(t)-\boldsymbol{x}_j^*(t)\right\|\\ &\quad+\sum_{j=1}^{n}\sum_{t=1}^{T}(\tilde{f}_{t,j}(\boldsymbol{x}_i^*(t))-\tilde{f}_{t,j}(\boldsymbol{x})).\end{aligned}$$

Note that, as the parameters are set to be $\eta_1=\cdots=\eta_n=\eta$, the last term in the right side is exactly the regret of the D-ODA algorithm applied to $\tilde{f}_{t,j}(\boldsymbol{x})$. In addition, using the results in Lemma 5, the sum of the first and the second terms is further bounded by $8nLDT^{3/4}$. Thus, we have

$$\begin{aligned}R_T(\boldsymbol{x}_i,\boldsymbol{x}) &\leq 8nLDT^{3/4}+R_T^a(\boldsymbol{x}_i,\boldsymbol{x})\\ &\leq 8nLDT^{3/4}+\frac{6\sqrt{n}+1-\sigma_2(P)}{2(1-\sigma_2(P))}\eta L^2T\\ &\quad+\frac{1}{\eta}D^2.\end{aligned}$$

Let $\eta=\frac{(1-\sigma_2(P))D}{2(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))LT^{3/4}}$. Then via a bit of analysis, we can verify that the choice of η_i satisfies the constraint required in Lemma 4

$$\eta_i(\frac{1+\sigma_2(P)}{1-\sigma_2(P)}\sqrt{n}+1)L\sqrt{h_{t+1,i}}\leq\sigma_{t,i}^2D^2.$$

The detailed verification is presented in the Appendix.

We thus finally obtain

$$\begin{aligned}R_T(\boldsymbol{x}_i,\boldsymbol{x}) &\leq 8nLDT^{3/4}\\ &\quad+\frac{6\sqrt{n}+1-\sigma_2(P)}{4(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))}LDT^{1/4}\\ &\quad+\frac{2(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))}{1-\sigma_2(P)}LDT^{3/4}.\end{aligned}$$

□

4. Experiments

To evaluate the performance of the proposed D-OCG algorithm, we conduct simulation experiments for a popular machine learning problem: multiclass classification.

4.1. Experimental Setup

Multiclass Classification In the distributed online learning setting, the problem is as follows. At each round $t=1,\cdots,T$, each learner i is presented with a data example $\boldsymbol{e}_i(t)\in\mathbb{R}^k$ which belongs to one of the classes $\mathcal{C}=\{1,\cdots,h\}$ and is required to generate a decision matrix $\boldsymbol{X}_i(t)=[\boldsymbol{x}_1^T;\cdots;\boldsymbol{x}_h^T]\in\mathbb{R}^{h\times k}$ that predicts the class label with $\arg\max_{\ell\in\mathcal{C}}\boldsymbol{x}_\ell^T\boldsymbol{e}_i(t)$. Then the adversary reveals the true class labels $y_i(t)$ and each learner i suffers a convex multivariate logistic loss

$$f_{t,i}(\boldsymbol{X}_i(t))=\log\Big(1+\sum_{\ell\neq y_i(t)}\exp(\boldsymbol{x}_\ell^T\boldsymbol{e}_i(t)-\boldsymbol{x}_{y_i(t)}^T\boldsymbol{e}_i(t))\Big).$$

The convex domain of the decision matrices is $\mathcal{K}=\{\boldsymbol{X}\in\mathbb{R}^{h\times k}|\,\|\boldsymbol{X}\|_{tr}\leq\tau\}$, where $\|\cdot\|_{tr}$ denotes the nuclear norm of matrices. In this case, the linear minimization required in each iteration of D-OCG amounts to compute a matrix's top singular vector, an operation that can be done in time near linear to the number of non-zeros in the matrix, whereas the projection onto $\mathcal{K}$ operation needed in traditional distributed online algorithms amounts to performing a full SVD, an $\mathcal{O}(hk\min(h,k))$ time operation that is much more expensive.

Datasets We use two multiclass datasets selected from the LIBSVM[2] repository with relatively large number of instances, which is summarized in Table 1.

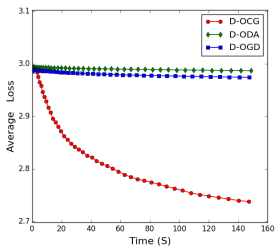
dataset	# features	# classes	# instances
news20	62,061	20	15,935
aloi	128	1,000	108,000

Table 1. Summary of the multiclass datasets

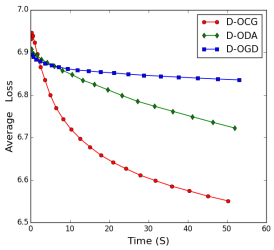
Network Topology To investigate the influence of network topology, we conduct our experiments on three types of graphs, which represent different levels of connectivity.

- *Complete graph.* This represents the highest level of connectivity in our experiments: all nodes are connected to each other.
- *Cycle graph.* This represents the lowest level of connectivity in our experiments: each node has only two immediate neighbors.
- *Watts-Strogatz.* This random graph generation technique (Watts & Strogatz, 1998) has two tunable pa-

[2]https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

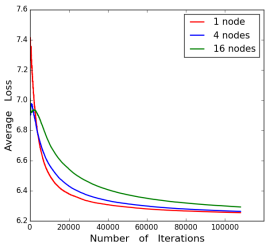


(a) news20

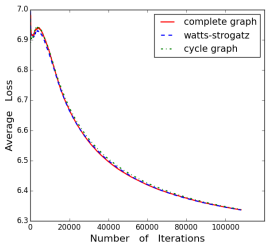


(b) aloi

Figure 1. Comparison of D-OCG, D-OGD and D-ODA on two multiclass datasets on a complete graph with 9 nodes



(a) size



(b) toplogy

Figure 2. (a): Comparison of D-OCG on graphs with different sizes; (b): Comparison of D-OCG on graphs with different topology and fixed 16 nodes (both on the aloi dataset)

rameters: the average degree of the graph k and the rewiring probability p. In general, the higher the rewiring probability, the better the connectivity of the graph (Colin et al., 2016). We tune the parameters $k = 4$ and $p = 0.3$ to achieve an intermediate level of connectivity in our experiments.

Compared Algorithms To evaluate the performance benefit of D-OCG over its counterparts with projection operation, we compare it with two classic algorithms: D-OGD (Yan et al., 2013) and D-ODA (Hosseini et al., 2013). To verify that performing online conditional gradient in the distributed setting does not lose much quality compared with that in the centralized setting, we also compare D-OCG with OCG, i.e. D-OCG with 1 node.

Parameter Settings We set most of the parameters in these algorithms as what their corresponding theories suggest. For instance, the parameters $\sigma_{t,i}$ in D-OCG are strictly set to be $\frac{1}{\sqrt{t}}$ and the learning rates in D-OGD are set to be the typical decaying sequence $\frac{1}{\sqrt{t}}$. We use the method utilized in (Duchi et al., 2012) to generate the doubly stochastic matrices and fix the nuclear norm bound τ to 50 throughout.

4.2. Experimental Results

We measure the running time of the D-OGD, D-ODA and D-OCG algorithms run on a complete graph with 9 nodes and see how fast the average losses decrease. From the results shown in Figure 1, we can clearly observe that D-OCG is significantly faster than both D-OGD and D-ODA, which illustrates the necessity and usefulness of using conditional gradient in distributed online learning.

We then investigate how the number of nodes affects the performance of D-OCG by running experiments on complete graphs with varying number of nodes. From the results shown in Figure 2(a), we can make the following two main observations. First, the average losses decrease more slowly on larger graphs than on smaller graphs, which nicely confirms our theoretical results. Second, D-OCG is able to yield comparable results to the centralized OCG.

We finally test the influence of network topology on the algorithm's performance. We run experiments on the aforementioned three types of graphs with 16 nodes using the aloi dataset. As shown in Figure 2(b), graphs with better connectivity lead to slightly faster convergence, which illustrates good agreement of empirical results with our theoretical predictions.

5. Conclusion

In this paper, we propose the distributed online conditional gradient algorithm for projection-free distributed online learning in networks. We give detailed analysis of the regret bound for the proposed algorithm, which depends on both the network size and the network topology. We evaluate the efficacy of the proposed algorithm on two real-world datasets for a multiclass classification task and find that it runs significantly faster than the counterpart algorithms with projection. The theoretical results regarding the regret bound for different graphs have also been verified.

Acknowledgements

This work is supported by National Program on Key Basic Research Project No. 2015CB352300 and National Natural Science Foundation of China Major Project No. U1611461. It is also supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative. We thank Zheng Xiong for helping constructing the networks and thank Wei Liu for his kind help in preparing the submission and the rebuttal. We finally acknowledge anonymous reviewers for their insightful comments on comparison and explanation of the regret bound.

References

Berman, Abraham and Plemmons, Robert J. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979.

Clarkson, Kenneth L. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):63, 2010.

Colin, Igor, Bellet, Aurelien, Salmon, Joseph, and Clémençon, Stéphan. Gossip dual averaging for decentralized optimization of pairwise functions. In *International Conference on Machine Learning*, pp. 1388–1396, 2016.

Duchi, John C, Agarwal, Alekh, and Wainwright, Martin J. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, 2012.

Dudík, Miroslav, Malick, Jérôme, et al. Lifted coordinate descent for learning with trace-norm regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 327–336, 2012.

Frank, Marguerite and Wolfe, Philip. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

Garber, Dan. Faster projection-free convex optimization over the spectrahedron. In *Advances In Neural Information Processing Systems*, pp. 874–882, 2016.

Garber, Dan and Hazan, Elad. A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26(3):1493–1528, 2016.

Harchaoui, Zaid, Juditsky, Anatoli, and Nemirovski, Arkadi. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015.

Hazan, Elad. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Hazan, Elad and Kale, Satyen. Projection-free online learning. In *International Conference on Machine Learning*, pp. 521–528, 2012.

Hazan, Elad and Luo, Haipeng. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pp. 235–243, 2016.

Hosseini, Saghar, Chapman, Airlie, and Mesbahi, Mehran. Online distributed optimization via dual averaging. In *IEEE Conference on Decision and Control*, pp. 1484–1489. IEEE, 2013.

Jaggi, Martin. Revisiting frank-wolfe: projection-free sparse convex optimization. In *International Conference on Machine Learning*, pp. 427–435, 2013.

Lee, Soomin, Nedic, Angelia, and Raginsky, Maxim. Decentralized online optimization with global objectives and local communication. *arXiv preprint arXiv:1508.07933*, 2015.

Nesterov, Yurii. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Ram, S Sundhar, Nedić, Angelia, and Veeravalli, Venugopal V. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications*, 147(3):516–545, 2010.

Sayed, Ali H et al. Adaptation, learning, and optimization over networks. *Foundations and Trends® in Machine Learning*, 7(4-5):311–801, 2014.

Watts, Duncan J and Strogatz, Steven H. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

Xiao, Lin. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

Yan, Feng, Sundaram, Shreyas, Vishwanathan, SVN, and Qi, Yuan. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25 (11):2483–2493, 2013.