Gradient-Free Method for Heavily Constrained Nonconvex Optimization

Wanli Shi [1 2] **Hongchang Gao** [3] **Bin Gu** [1 2]

Abstract

Zeroth-order (ZO) method has been shown to be a powerful method for solving the optimization problem where explicit expression of the gradients is difficult or infeasible to obtain. Recently, due to the practical value of the constrained problems, a lot of ZO Frank-Wolfe or projected ZO methods have been proposed. However, in many applications, we may have a very large number of nonconvex white/black-box constraints, which makes the existing zeroth-order methods extremely inefficient (or even not working) since they need to inquire function value of all the constraints and project the solution to the complicated feasible set. In this paper, to solve the nonconvex problem with a large number of white/black-box constraints, we proposed a doubly stochastic zeroth-order gradient method (DSZOG) with momentum method and adaptive step size. Theoretically, we prove DSZOG can converge to the ϵ-stationary point of the constrained problem. Experimental results in two applications demonstrate the superiority of our method in terms of training time and accuracy compared with other ZO methods for the constrained problem.

1. Introduction

Zeroth-order (gradient-free) method is a powerful method for solving the optimization problem where explicit expression of the gradients are difficult or infeasible to obtain, such as bandit feedback analysis (Agarwal et al., 2010), reinforcement learning (Choromanski et al., 2018), and adversarial attacks on black-box deep neural networks (Chen et al., 2017; Liu et al., 2018b). Recently, more and more zeroth-order gradient algorithms have been proposed and achieved great success, such as (Ghadimi & Lan, 2013; Wang et al., 2018; Gu et al., 2016; Liu et al., 2018b; Huang et al., 2020a; Gu et al., 2021b; Wei et al., 2021; Gu et al., 2021a).

Due to several motivating applications, the study of the zeroth-order methods in constrained optimization has gained great attention. For example, ZOSCGD(Balasubramanian & Ghadimi, 2018) uses the zeroth-order method to approximate the unbiased stochastic gradient of the objective, and then uses the Frank-Wolfe framework to update the parameters. (Gao & Huang, 2020; Huang et al., 2020b) apply the variance reduction technique (Fang et al., 2018; Nguyen et al., 2017) or momentum method in ZOSCGD and obtain a better convergence performance. In addition, ZOSPGD (Liu et al., 2018c) uses the zeroth-order gradient to update the parameters and then projects the parameters onto the feasible subset. The variance reduction and momentum methods are also used to obtain a better performance (Huang et al., 2020a). We have summarized several representative zeroth-order methods for constrained optimization in Table 1.

However, all these methods are not scalable for the problems with a large number of constraints. On the one hand, they all need to evaluate the values of all the constraints in each iteration. On the other hand, the projected gradient methods and the Frank-Wolfe methods need to solve a subproblem in each iteration. These makes the existing methods time-consuming to find a feasible point. What's worse, all these methods need the constraints to be convex white-box functions. However, in many real-world applications, the constraints could be nonconvex or black-box functions, which means existing methods are extremely inefficient or even not working. Therefore, how to effectively solve the nonconvex constrained problem with a large number of nonconvex/convex white/black-box constraints, which is denoted as heavily constrained problem, by using the ZO method is still an open problem.

In this paper, to solve the heavily constrained nonconvex optimization problem efficiently, we propose a new ZO algorithms called doubly stochastic zeroth-order gradient method (DSZOG). Specifically, we give a probability distribution over all the constraints and rewrite the original problem as a nonconvex-strongly-concave minimax problem (Lin et al., 2020; Wang et al., 2020; Huang et al., 2020a; Guo et al., 2021) by using the penalty method. We sample a batch of training points uniformly and a batch of constraints

[1]Nanjing University of Information Science and Technology, Jiangsu, China [2]MBZUAI, Abu Dhabi, UAE [3]Department of Computer and Information Sciences, Temple University, PA, USA. Correspondence to: Bin Gu <jsgubin@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Table 1: Representative zeroth order methods for constrained optimization problems, where N/C means nonconvex/convex, W/B means white/black-box function, and the last column shows the size of the constraints.

Framework	Algorthm	Reference	Objective	Constraints	Size
Frank-Wolfe	ZOSCGD	(Balasubramanian & Ghadimi, 2018)	N/C	C W	Small
	FZFW FZCGS FCGS	(Gao & Huang, 2020)	N/C	C W	Small
	Acc-SZOFW Acc-SZOFW*	(Huang et al., 2020b)	N/C	C W	Small
Projected	ZOPSGD	(Liu et al., 2018c)	N/C	C W	Small
	AccZOMDA	(Huang et al., 2020a)	N/C	C W	Small
Penalty	DSZOG	Ours	N/C	N/C W/B	Large

according to the distribution to calculate the zeroth-order gradient of the penalty function w.r.t model parameters and then sample a batch of constraints uniformly to calculate the stochastic gradient of penalty function w.r.t the probability distribution. Then, gradient descent and projected gradient ascent can be used to update model parameters and probability distribution, respectively. In addition, we also use the exponential moving average (EMA) method and adaptive stepsize (Guo et al., 2021; Huang et al., 2020a), which benefits our method from the variance reduction and adaptive convergence. Theoretically, we prove DSZOG can converge to the ϵ-stationary point of the constrained problem. Experimental results in two applications demonstrate the superiority of our method in terms of training time and accuracy compared with other ZO methods for constrained problem.

Contributions. We summarized the main contributions of this paper as follows:

1. We propose a doubly stochastic zeroth-order gradient method to solve the heavily constrained nonconvex problem. By introducing a stochastic layer into the constraints, our method is scalable and efficient for the heavily constrained nonconvex problem.

2. By using the exponential moving average method and adaptive stepsize, our method enjoys the benefits of variance reduction and adaptive convergence.

3. We prove DSZOG can converge to the ϵ-stationary point of the constrained problem. Experimental results also demonstrate the superiority of our methods in terms of accuracy and training time.

2. Related Works

2.1. Zeroth-Order Methods

Zeroth-order methods are powerful methods to solve several machine learning problems, where the explicit gradients are difficult or infeasible to obtain. Based on Gaussian smoothing method, (Ghadimi & Lan, 2013; Duchi et al., 2015; Nesterov & Spokoiny, 2017; Gu et al., 2016; 2021b; Wei et al., 2021; Gu et al., 2021a) propose several zeroth-order method which only needs function values to estimate the gradients. To deal with nonsmooth optimization problem, some zeroth-order proximal gradient methods (Ghadimi et al., 2016; Ji et al., 2019) and ADMM methods (Gao et al., 2018; Liu et al., 2018a) have been proposed. In addition, to solve the constrained optimization problems, the zeroth-order projection method (Liu et al., 2018c) and the zeroth-order Frank-Wolfe methods (Balasubramanian & Ghadimi, 2018; Chen et al., 2020) have been proposed. More recently, based on the variance reduced techniques, some accelerated zeroth-order stochastic methods have been proposed. (Gu et al., 2021b) proposed a new framework to reduce the query complexities of zeroth-order gradient methods for convex and nonconvex objectives. In addition, this new framework can be used in various ZO method and has a better convergence performance.

2.2. Variance Reduction and Momentum Methods

To accelerate stochastic gradient descent method, variance reduction methods such as SAG (Roux et al., 2012), SAGA (Defazio et al., 2014), SVRG (Johnson & Zhang, 2013), SARAH (Nguyen et al., 2017) have been proposed. Recently, due to the widely existence of nonconvex optimization, several variance reduction methods for nonconvex optimization have been proposed (Allen-Zhu, 2017; Lei et al., 2017; Fang et al., 2018; Wang et al., 2019; Zhou et al., 2018). Another method to accelerate the stochastic gradient method is to use momentum-based method. For convex

and nonconvex optimization problem, various momentum-based methods have been proposed, e.g. APCG (Lin et al., 2014), Katyusha (Allen-Zhu, 2017), STORM (Cutkosky & Orabona, 2019), NIGHT (Cutkosky & Mehta, 2020), Hybrid-SGD (Tran-Dinh et al., 2021), etc. Due to the superiority of variance reduction and momentum methods, they have been widely used in zeroth-order gradient methods and have achieved great success.

3. Preliminaries

3.1. Problem Setting

In this paper, we consider the following nonconvex constrained problem,

$$\min_{\boldsymbol{w}} \ f_0(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{w}), \qquad (1)$$
$$s.t.\ f_j(\boldsymbol{w}) \leq 0,\ j = 1, \cdots, m,$$

where $\boldsymbol{w} \in \mathbb{R}^d$ is the optimization variable, $\{\ell_i(\boldsymbol{w})\}_{i=1}^n$ are n component functions. In addition, $f_0 : \mathbb{R}^d \mapsto \mathbb{R}$ is a nonconvex and black-box function. $f_j : \mathbb{R}^d \mapsto \mathbb{R}$, $(j = 1, \cdots, m)$, is nonconvex/convex and white/black-box function. We can denote such problem as heavily constrained problem.

3.2. Reformulate the Constrained Problem

To solve the constrained problem, the penalty method is one of the main approaches and has achieved great success (Clarkson et al., 2012; Cotter et al., 2016; Shi & Gu, 2021). Specifically, the penalty method reformulates the problem by adding a new term onto the objective to penalize the constraints and the solves the new problem to find a KKT point. Based on the penalty method, we reformulate the constrained optimization problem 1 as the following minimax problem over a probability distribution (Clarkson et al., 2012; Cotter et al., 2016)

$$\min_{\boldsymbol{w}} \max_{\boldsymbol{p}\in\Delta^m} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p}) = f_0(\boldsymbol{w}) + \beta\varphi(\boldsymbol{w}, \boldsymbol{p}) - \frac{\lambda}{2}\|\boldsymbol{p}\|_2^2, \quad (2)$$

where $\beta > 0$, $\lambda > 0$, $\varphi(\boldsymbol{w}, \boldsymbol{p}) = \sum_{j=1}^m p_j\phi_j(\boldsymbol{w})$, $\phi_j(\boldsymbol{w}) = (\max\{f_j(\boldsymbol{w}), 0\})^2$ is the penalty function on f_j, $\Delta^m := \{\boldsymbol{p} | \sum_{j=1}^d p_j = 1, 0 \leq p_j \leq 1, \forall j \in [d]\}$ is the m-dimensional simplex and $\boldsymbol{p} = [p_1, \cdots, p_m] \in \Delta^m$. We add an additional term term $-\frac{1}{2}\|\boldsymbol{p}\|_2^2$ in problem 2 to ensure $\mathcal{L}$ is strongly concave on $\boldsymbol{p}$.

Since we can only obtain the values of the objective and constraints, stochastic zeroth-order gradient method is one of the effective ways to solve this problem. However, calculating the stochastic zeroth-order gradient of $\mathcal{L}$ needs to inquire the function values of all the constraints, which has a very high time complexity if m is very large. This make it time-consuming.

4. Proposed Method

4.1. Doubly Stochastic Zeroth-order Gradient Method

To solve problem 2 efficiently, we introduce the another stochastic layer to the constraints. Specifically, since the minimax problem 2 contains two finite sums, i.e., $f_0(\boldsymbol{w}) = 1/n\sum_{i=1}^n \ell_i(\boldsymbol{w})$ and $\varphi(\boldsymbol{w}, \boldsymbol{p}) = \sum_{j=1}^m p_j\phi_j(\boldsymbol{w})$, we can calculate their stochastic zeroth-order gradient, respectively, and then combine these two gradient to obtain the stochastic zeroth-order gradient of $\mathcal{L}$.

We can calculate the stochastic zeroth-order gradient of $f_0(\boldsymbol{w})$ and $\varphi(\boldsymbol{w}, \boldsymbol{p})$ as follows,

$$G_\mu^f(\boldsymbol{w}_t, \ell_i, \boldsymbol{u}) = \frac{\ell_i(\boldsymbol{w}_t + \mu\boldsymbol{u}) - \ell_i(\boldsymbol{w}_t)}{\mu}\boldsymbol{u}, \qquad (3)$$

$$G_\mu^\varphi(\boldsymbol{w}_t, \boldsymbol{p}, f_j, \boldsymbol{u}) = \frac{\phi_j(\boldsymbol{w}_t + \mu\boldsymbol{u}) - \phi_j(\boldsymbol{w}_t)}{\mu}\boldsymbol{u}, \qquad (4)$$

by sampling ℓ_i uniformly, and f_j according to $\boldsymbol{p}$, where $\mu > 0$ and $\boldsymbol{u} \sim \mathcal{N}(0, \mathbf{1}_d)$. Then, combining these two terms, we can obtain the stochastic zeroth-order gradient of $\mathcal{L}$ w.r.t. $\boldsymbol{w}$ as follows,

$$\begin{aligned} &G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_i, f_j, \boldsymbol{u}) \\ =&G_\mu^f(\boldsymbol{w}_t, \ell_i, \boldsymbol{u}) + \beta G_\mu^\varphi(\boldsymbol{w}_t, \boldsymbol{p}_t, f_j, \boldsymbol{u}). \end{aligned} \qquad (5)$$

To reduce the variance, we can sample a batch of ℓ_i, f_j and $\boldsymbol{u}_k$ to calculate the zeroth-order gradient. Given $q > 0$, $\mathcal{M}_1 \subseteq [n]$ and $\mathcal{M}_2 \sim \boldsymbol{p} \subseteq [m]$, we have

$$\begin{aligned} &G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \\ =&\frac{1}{q|\mathcal{M}_1|}\sum_{i\in\mathcal{M}_1}\sum_{k=1}^{q} G_\mu^f(\boldsymbol{w}_t, \ell_i, \boldsymbol{u}_k) \\ &+ \frac{\beta}{q|\mathcal{M}_2|}\sum_{j\in\mathcal{M}_2}\sum_{k=1}^{q} G_\mu^\varphi(\boldsymbol{w}_t, \boldsymbol{p}_t, f_j, \boldsymbol{u}_k), \end{aligned} \qquad (6)$$

Then, the gradient descent can be used to update $\boldsymbol{w}$ by using the following rule

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_{\boldsymbol{w}} G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}). \qquad (7)$$

Then, in each iteration, we randomly sample a constraint $f_j(\boldsymbol{w})$ to calculate the stochastic gradient of $\mathcal{L}$ w.r.t. $\boldsymbol{p}$ by using

$$H(\boldsymbol{w}_t, \boldsymbol{p}_t, f_j) = \beta m \boldsymbol{e}_j\phi_j(\boldsymbol{w}_t) - \lambda\boldsymbol{p}_t, \qquad (8)$$

where $\boldsymbol{e}_j$ is the jth m-dimensional standard unit basis vector. Mini-batch can be also used to reduce variance. Assume

we have the randomly sampled index set $\mathcal{M}_3 \subseteq [m]$, the mini-batch gradient of $\mathcal{L}$ w.r.t $\boldsymbol{p}$ becomes

$$H(\boldsymbol{w}_t, \boldsymbol{p}_t, f_{\mathcal{M}_3}) = \frac{\beta m}{|\mathcal{M}_3|} \sum_{j \in \mathcal{M}_3} \boldsymbol{e}_j \phi_j(\boldsymbol{w}_t) - \lambda \boldsymbol{p}_t. \quad (9)$$

Then we can perform gradient ascent by using the following rules,

$$\boldsymbol{p}_{t+1} = \mathcal{P}_{\Delta^m}(\boldsymbol{p}_t + \eta_{\boldsymbol{p}} H(\boldsymbol{w}_t, \boldsymbol{p}_t, f_{\mathcal{M}_3}), \quad (10)$$

to update $\boldsymbol{p}$, where $\mathcal{P}_{\Delta^m}(\cdot)$ denotes the projection onto Δ^m and is easy to calculate.

Note that since m and n are sufficient large in this problem, $G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})$ and $H(\boldsymbol{w}_t, \boldsymbol{p}_t, f_{\mathcal{M}_3})$ can be viewed as the unbiased estimation of the gradients of $\mathcal{L}$ w.r.t $\boldsymbol{w}$ and $\boldsymbol{p}$, respectively.

4.2. Momentum and Adaptive Step Size

To further improve our method, we use exponential moving average (EMA) method (Wang et al., 2017; Liu et al., 2020; Cutkosky & Mehta, 2020; Guo et al., 2021) and adaptive stepsize. We use the following exponential moving average (EMA) method on the zeroth-order and first-order gradient to smooth out short-term fluctuations, highlight longer-term trends and reduce the variance of stochastic gradient (Wang et al., 2017; Guo et al., 2021)

$$\boldsymbol{z}_{\boldsymbol{w}}^{t+1} = (1-b)\boldsymbol{z}_{\boldsymbol{w}}^t + bG_\mu^{\mathcal{L}}(\boldsymbol{w}_{t+1}, \boldsymbol{p}_{t+1}, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}), \quad (11)$$

$$\boldsymbol{z}_{\boldsymbol{p}}^{t+1} = (1-b)\boldsymbol{z}_{\boldsymbol{p}}^t + bH(\boldsymbol{w}_{t+1}, \boldsymbol{p}_{t+1}, f_{\mathcal{M}_3}), \quad (12)$$

where $0 < b < 1$, $\boldsymbol{z}_{\boldsymbol{w}}^1 = G_\mu^{\mathcal{L}}(\boldsymbol{w}_1, \boldsymbol{p}_1, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})$ and $\boldsymbol{z}_{\boldsymbol{p}}^1 = H(\boldsymbol{w}_1, \boldsymbol{p}_1, f_{\mathcal{M}_3})$. Here, $H(\boldsymbol{w}_{t+1}, \boldsymbol{p}_{t+1}, f_{\mathcal{M}_3})$ is calculated on the intermediate point $\boldsymbol{p}_{t+1} = (1-a)\boldsymbol{p}_t + a\hat{\boldsymbol{p}}_{t+1}$, where $0 < a < 1$ and $\hat{\boldsymbol{p}}_{t+1}$ is the distribution after updating and projecting onto the Δ^m.

Then we use adaptive stepsizes to update $\boldsymbol{w}$ and $\boldsymbol{p}$. Specifically, we ensure the stepsizes are proportional to $1/(\sqrt{\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2} + c)$ and $1/(\sqrt{\|\boldsymbol{z}_{\boldsymbol{p}}^t\|_2} + c)$ (Liu et al., 2020; Guo et al., 2021), where $c > 0$ is a small constant used to prevent the denominator from becoming 0. Therefore, the update rules of $\boldsymbol{w}$ and $\boldsymbol{p}$ become

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_{\boldsymbol{w}} \frac{\boldsymbol{z}_{\boldsymbol{w}}^t}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2} + c}, \quad (13)$$

$$\hat{\boldsymbol{p}}_{t+1} = \mathcal{P}_{\Delta^m}(\boldsymbol{p}_t + \eta_{\boldsymbol{p}} \frac{\boldsymbol{z}_{\boldsymbol{p}}^t}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{p}}^t\|_2} + c}). \quad (14)$$

These two key components of our method, *i.e.,* extrapolation moving average and adaptive stepsize from the root norm of the momentum estimate, make our method enjoy two noticeable benefits: variance reduction of momentum estimate and adaptive convergence. The whole algorithm is presented in Algorithm 1. Since there exist two sources of randomness, we call our method doubly stochastic zeroth-order gradient method (DSZOG). Note that different from the algorithm in (Guo et al., 2021), we use the adaptive step size method in both updating $\boldsymbol{w}$ and $\boldsymbol{p}$.

Algorithm 1 Doubly Stochastic Zeroth-order Gradient (DSZOG).

Input: T, $|\mathcal{M}_1|, |\mathcal{M}_2|, |\mathcal{M}_3|$, $\beta \geq 1$, q, μ, $\lambda = 1e-6$, $b \in (0,1)$, $c = 1e-8$, $a \in (0,1)$, η_w and $\eta_{\boldsymbol{p}}$.
Output: $\boldsymbol{w}_T$.

```
Initialize w_1.
Initialize p_1 = p*(w_1) by solving the strongly concave problem.
Initialize z_w^1 = G_μ^L(w_1, p_1, ℓ_{M_1}, f_{M_2}, u_[q]) and z_p^1 = H(w_1, p_1, f_{M_3}).
for t = 1, ..., T do
    w_{t+1} = w_t − η_w z_w^t / (sqrt(||z_w^t||_2) + c).
    p̂_{t+1} = P_{Δ^m}(p_t + η_p z_p^t / (sqrt(||z_p^t||_2) + c)).
    p_{t+1} = (1 − a)p_t + a p̂_{t+1}.
    Randomly sample u_1, ..., u_q ~ N(0, 1_d).
    Randomly sample a index set M_1 ⊆ [n] of ℓ_i.
    Sample a constraint index set M_2 ~ p_{t+1} ⊆ [m].
    Randomly sample a constraint index set M_3.
    Calculate G_μ^L(w_{t+1}, p_{t+1}, ℓ_{M_1}, f_{M_2}, u_[q]) = 1/(q|M_1|) Σ_{i∈M_1} Σ_{k=1}^q G_μ^f(w_{t+1}, ℓ_i, u_k) + 1/(q|M_2|) Σ_{j∈M_2} Σ_{k=1}^q G_μ^φ(w_{t+1}, p_{t+1}, f_j, u_k).
    Calculate H(w_{t+1}, p_{t+1}, f_{M_3}) = βm/|M_3| Σ_{j∈M_3} e_j φ_j(w_{t+1}) − λ p_{t+1}.
    z_w^{t+1} = (1 − b) z_w^t + b G_μ^L(w_{t+1}, p_{t+1}, ℓ_{M_1}, f_{M_2}, u_[q]).
    z_p^{t+1} = (1 − b) z_p^t + b H(w_{t+1}, p_{t+1}, f_{M_3}).
end for
```

5. Convergence Analysis

In this section, we discuss the convergence performance of our methods. The detailed proofs are given in our appendix.

5.1. Stationary point

In this subsection, we first give the assumption about $\mathcal{L}$ and then give the definitions of the stationary point.

Assumption 5.1. The objective function $\mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$ has the

following properties:

1. $\mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$ is continuously differentiable in $\boldsymbol{w}$ and $\boldsymbol{p}$. $\mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$ is nonconvex with respect to $\boldsymbol{w}$, and $\mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$ is τ-strongly concave with respect to $\boldsymbol{p}$.
2. The function $g(\boldsymbol{w}) := \max_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$ is lower bounded, and $g(\boldsymbol{w})$ is L_g-Lipschitz continuous.
3. When viewed as a function in $\mathbb{R}^{d+m}$, $\mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$ is L-gradient Lipschitz, ($L > 0$), such that $\|\nabla \mathcal{L}(\boldsymbol{w}_1, \boldsymbol{p}_1) - \nabla \mathcal{L}(\boldsymbol{w}_2, \boldsymbol{p}_2)\|_2 \leq L\|(\boldsymbol{w}_1, \boldsymbol{p}_1) - (\boldsymbol{w}_2, \boldsymbol{p}_2)\|_2$.

This assumption is widely used in the convergence analysis of minimax problems (Wang et al., 2020; Huang et al., 2020a). The first condition is used to detail the structure of $\mathcal{L}$ and the second condition is used to make the optimization problem well defined , and the third condition places a restriction on the degree of smoothness to be satisfied by the objective function.

Then, we discuss the definitions of stationary points and their relationships. For a general nonconvex constrained optimization problem, the stationary point (Lin et al., 2019) is defined as follows,

Definition 5.2. $\boldsymbol{w}^*$ is said to be the stationary point of problem (1), if the following conditions holds,

$$\nabla_{\boldsymbol{w}} f_0(\boldsymbol{w}^*) + \sum_{j=1}^{m} \alpha_j^* \nabla_{\boldsymbol{w}} f_j(\boldsymbol{w}^*) = \boldsymbol{0}, \tag{15}$$

$$f_j(\boldsymbol{w}^*) \leq 0, \tag{16}$$

$$\alpha_j^* f_j(\boldsymbol{w}^*) = 0, \quad \forall i \in \{1, \cdots, m\}, \tag{17}$$

where $\boldsymbol{\alpha}^* = [\alpha_1, \cdots, \alpha_m]_t$ denotes the Lagrangian multiplier and $\alpha_j \geq 0, \forall i = 1, \cdots, m$.

However, it is hard to compute a solution that satisfies the above conditions exactly (Lin et al., 2019). Therefore, finding the following ϵ-stationary point (Lin et al., 2019) is more practicable,

Definition 5.3. (ϵ-stationary) $\boldsymbol{w}^*$ is said to be the ϵ-stationary point of problem (1), if there exists a vector $\boldsymbol{\alpha}^* \geq \boldsymbol{0}$, such that the following conditions hold,

$$\|\nabla_{\boldsymbol{w}} f_0(\boldsymbol{w}^*) + \sum_{j=1}^{m} \alpha_j^* \nabla_{\boldsymbol{w}} f_j(\boldsymbol{w}^*)\|_2^2 \leq \epsilon_1^2, \tag{18}$$

$$\sum_{j=1}^{m} (\max\{f_j(\boldsymbol{w}^*), 0\})^2 \leq \epsilon_2^2, \tag{19}$$

$$\sum_{j=1}^{m} (\alpha_j f_j(\boldsymbol{w}^*))^2 \leq \epsilon_3^2. \tag{20}$$

Since we reformulate the constrained problem as a minimax problem, here we give the definition of the approximation stationary point of the minimax problem and then show its relationship with Definition 5.3. According to (Wang et al., 2020), we have the following definition,

Definition 5.4. A point $(\boldsymbol{w}^*, \boldsymbol{p}^*)$ is called the ϵ-stationary point of problem $\min_{\boldsymbol{w}} \max_{\boldsymbol{p} \in \Delta^m} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$ if it satisfies the conditions: $\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})\|_2^2 \leq \epsilon^2$ and $\|\nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})\|_2^2 \leq \epsilon^2$.

In addition, we have the following Proposition between definition 5.3 and definition 5.4.

Proposition 5.5. *If Assumption 5.1 holds,* $\sqrt{\frac{2m\epsilon^2 + 2m^2\lambda^2}{\beta^2}} \leq \epsilon_2^2$ *and* $(\boldsymbol{w}^*, \boldsymbol{p}^*)$ *is the* ϵ*-stationary point defined in Definition 5.4 of the problem* $\min_{\boldsymbol{w}} \max_{\boldsymbol{p} \in \Delta^m} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha})$*, then* $\boldsymbol{w}^*$ *is the* ϵ*-stationary point defined in Definition 5.3 of the constrained problem 1.*

As proposed in (Wang et al., 2020), the minimax problem 2 is equivalent to the following minimization problem:

$$\min_{\boldsymbol{w}} \left\{ g(\boldsymbol{w}) := \max_{\boldsymbol{p} \in \Delta^m} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p}) = \mathcal{L}(\boldsymbol{w}, \boldsymbol{p}^*(\boldsymbol{w})) \right\}, \tag{21}$$

where $\boldsymbol{p}^*(\boldsymbol{w}) = \arg\max_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$. Here, we give stationary point the minimization problem 21 and its relationship with Definition 5.4 as follows,

Definition 5.6. We call $\boldsymbol{w}^*$ an ϵ-stationary point of a differentiable function $g(\boldsymbol{w})$, if $\|\nabla g(\boldsymbol{w}^*)\|_2 \leq \epsilon$.

Proposition 5.7. *Under Assumption 5.1, if a point* $\boldsymbol{w}'$ *is an* ϵ*-stationary point in terms of Definition 5.6, then an* ϵ*-stationary point* $\boldsymbol{w}^*, \boldsymbol{p}^*$ *in terms of Definition 5.4 can be obtained.*

Remark 5.8. According to Proposition 5.5 and Proposition 5.7, we have that once we find the ϵ-stationary point in terms of Definition 5.6, then we can get the ϵ-stationary point in terms of Definition 5.3.

5.2. Convergence Rate of the Accelerated Method

In this subsection, we discuss the convergence performance of our algorithms. Here, we give several assumptions used in our analysis.

Assumption 5.9. We have $c_{1,l} \leq \frac{1}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2 + c}} \leq c_{1,u}$ and $c_{2,l} \leq \frac{1}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{p}}^t\|_2 + c}} \leq c_{2,u}$, where c is a constant.

This assumption is used to bound the step size scaling factor which is widely used in (Huang et al., 2021; Guo et al., 2021; Huang & Huang, 2021).

Assumption 5.10. For any $\boldsymbol{w} \in \mathbb{R}^d$, the following proper-

ties holds,

$$\mathbb{E}[G_\mu^{\mathcal{L}}(\boldsymbol{w},\boldsymbol{p},\ell_{\mathcal{M}_1},f_{\mathcal{M}_2},\boldsymbol{u}_{[q]})]=\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w},\boldsymbol{p}),$$
$$\mathbb{E}[H(\boldsymbol{w},\boldsymbol{p},f_{\mathcal{M}_3})]=\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w},\boldsymbol{p}),$$
$$\mathbb{E}[\|G_\mu^{\mathcal{L}}(\boldsymbol{w}_1,\boldsymbol{p}_1,\ell_{\mathcal{M}_1},f_{\mathcal{M}_2},\boldsymbol{u}_{[q]})-\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)\|_2]\le\sigma_1^2,$$
$$\mathbb{E}[\|H(\boldsymbol{w},\boldsymbol{p},f_{\mathcal{M}_3})-\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)\|_2]\le\sigma_2^2.$$

.

This assumptions is used to show our estimator is the unbiased estimation. Based on above assumptions, we can derive the following lemmas which are useful in our convergence analysis.

Lemma 5.11. ***(Descent in the function value.)*** *Under Assumptions 5.1 and 5.9, if* $\eta_w L\le\frac{c_{1,l}}{2c_{1,u}^2}$, *we have*

$$\begin{aligned}&g(\boldsymbol{w}_{t+1})\\ \le&g(\boldsymbol{w}_t)-\frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2-\frac{\eta_w c_{1,l}}{4}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2\\&+\eta_w c_{1,l}\frac{\mu^2L^2(d+3)^3}{2}+\eta_w c_{1,l}L^2\|\boldsymbol{p}^*(\boldsymbol{w}_t)-\boldsymbol{p}_t\|_2^2\\&+\eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2.\end{aligned}$$

Lemma 5.12. ***(Descent in the iterates of the probability.)*** *Under Assumptions 5.1, 5.9 and 5.10, if* $a\le 1$ *and* $\eta_p\le\frac{1}{3c_{2.l}L}$, *we have*

$$\begin{aligned}&\|\boldsymbol{p}_{t+1}-\boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2\\ \le&-\frac{1}{4a}\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_2^2+\frac{8a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2\\&+(1-\frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}^*(\boldsymbol{w}_t)-\boldsymbol{p}_t\|_2^2\\&+\frac{8L_g^2}{\tau a\eta_p c_{2,l}}\|\boldsymbol{w}_t-\boldsymbol{w}_{t+1}\|_2^2.\end{aligned}$$

Lemma 5.13. ***(Descent in the gradient estimation error.)*** *Under Assumptions 5.1, 5.9 and 5.10, if* $b\in(0,1)$, *we have*

$$\begin{aligned}&\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\boldsymbol{z}_{\boldsymbol{p}}^{t+1}\|_2^2]\\ \le&(1-b)\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2]+\frac{1}{b}L^2[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2^2\\&+\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|_2^2]+b^2\sigma_2^2,\\&\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\boldsymbol{z}_{\boldsymbol{w}}^{t+1}\|_2^2]\\ \le&(1-b)\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2]+\frac{1}{b}L^2[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2^2\\&+\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|_2^2]+b^2\sigma_1^2.\end{aligned}$$

Then, following the framework in (Guo et al., 2021; Wang et al., 2018; Huang et al., 2020a) and utilizing the above lemmas, we have the following theorem,

Theorem 5.14. *Under Assumptions 5.1, 5.9 and 5.10, if* $a\in(0,1]$, $,\boldsymbol{p}^*(w_1)=\boldsymbol{p}_1$, $\boldsymbol{z}_{\boldsymbol{p}}^1=H(\boldsymbol{w}_t,\boldsymbol{p}_t,f_{\mathcal{M}_3})$, $\boldsymbol{z}_{\boldsymbol{w}}^1=G_\mu^{\mathcal{L}}(\boldsymbol{w}_t,\boldsymbol{p}_t,\ell_{\mathcal{M}_1},f_{\mathcal{M}_2},\boldsymbol{u}_{[q]})$, $0<\eta_p\le\min\{\frac{1}{3c_{2.l}L},\frac{b^2}{\tau a^2c_{2,l}},\frac{\tau b^2}{32L^2a^2c_{2,l}},1\}$, $0<\eta_w^2\le\min\{\frac{c_{2,l}^2}{4Lc_{1,u}^4},\frac{b^2}{4c_{1,u}^2L^2},\frac{\tau^2a^2\eta_p^2c_{2,l}^2}{128L_g^2L^2c_{1,u}},\frac{\tau^2b^2}{128L^4c_{1,u}^2},1\}$, $\mu\le\frac{\epsilon}{L(d+3)^{3/2}}$, $0<b\le\min\{\frac{\epsilon^2}{2\sigma_1^2},\frac{\tau^2\epsilon^2}{64\sigma_2^2L^2},1\}$ *and* $T\ge\max\{\frac{2(g(\boldsymbol{w}_1)-g(\boldsymbol{w}_T))}{\epsilon^2\eta_w c_{1,l}},\frac{2\sigma_1^2}{\epsilon^2b},\frac{64\sigma_2^2L^2}{\epsilon^2\tau^2b}\}$, *we have*

$$\frac{1}{T}\mathbb{E}[\sum_{t=1}^T\|\nabla g(\boldsymbol{w}_t)\|_2^2]\le\epsilon^2. \tag{22}$$

Remark 5.15. By choosing $b=\mathcal{O}(\epsilon^2/\kappa^2)$, $\eta_{\boldsymbol{w}}=\mathcal{O}(\epsilon^2/\kappa^6)$, $\eta_{\boldsymbol{p}}$=$\mathcal{O}(\epsilon^2/\kappa^4)$ and $T=\mathcal{O}(\kappa^6/\epsilon^4)$, our proposed DSZOG can converge to the ϵ-stationary point defined in Definition 5.6. Then, based on Proposition 5.5 and Proposition 5.7, we can derive that our method can converge to the ϵ-stationary point of the original constrained problem (1) defined in Definition 5.3.

6. Experiments

6.1. Experimental Setup

In this subsection, we summarized the baselines used in our experiments as follows,

1. **ZOPSGD**(Liu et al., 2018c). In each iteration, ZOPSGD calculates the stochastic zeroth-order gradient of f_0 to update the parameters and then solves a constrained quadratic problem to project the solution into the feasible set.

2. **ZOSCGD**(Balasubramanian & Ghadimi, 2018). In each iteration, ZOSCGD calculates the stochastic zeroth-order gradient of f_0 and then uses the conditional gradient method to update the parameters by solving a constrained linear problem.

3. **AccZOMDA**(Huang et al., 2020a). In each iteration, AccZOMDA uses the momentum-based variance reduce technique of STORM (Cutkosky & Orabona, 2019) to estimate the stochastic zeroth-order gradients, and then solves a constrained quadratic problem to project the solution into the feasible set.

4. **AccSZOFW**(Huang et al., 2020b).In each iteration, AccSZOFW uses the variance reduced technique of SPIDER (Fang et al., 2018) to calculate the stochastic zeroth-order gradient of f_0 and then uses the conditional gradient method to update the parameters by solving a constrained linear problem.

Table 2: Test accuracy (%) of all the methods in classification with pairwise constraints.

Data	DSZOG	ZOSCGD	ZOPSGD	AccZOMDA	AccSZOFW
a9a	**75.90** $\pm$ 0.26	75.35 $\pm$ 0.13	75.37 $\pm$ 0.19	75.52 $\pm$ 0.21	75.22 $\pm$ 0.12
w8a	**89.94** $\pm$ 0.28	83.53 $\pm$ 0.58	89.02 $\pm$ 0.97	89.14 $\pm$ 0.23	89.34 $\pm$ 0.33
gen	**82.33** $\pm$ 0.76	66.33 $\pm$ 0.07	66.83 $\pm$ 0.57	72.84 $\pm$ 0.45	72.03 $\pm$ 0.23
svm	**79.56** $\pm$ 0.49	71.21 $\pm$ 0.57	78.63 $\pm$ 0.26	79.01 $\pm$ 0.21	71.88 $\pm$ 0.34

Table 3: Test accuracy (%) of all the methods in classification with fairness constraints.

Data	DSZOG	ZOSCGD	ZOPSGD	AccZOMDA	AccSZOFW
D1	**87.33** $\pm$ 0.38	51.08 $\pm$ 0.57	59.16 $\pm$ 0.37	66.33 $\pm$ 0.19	55.23 $\pm$ 0.46
D2	**84.75** $\pm$ 0.25	69.70 $\pm$ 0.24	68.00 $\pm$ 0.54	69.55 $\pm$ 0.29	76.13 $\pm$ 0.45
D3	**83.58** $\pm$ 0.14	66.33 $\pm$ 0.30	66.84 $\pm$ 0.57	66.35 $\pm$ 0.45	60.47 $\pm$ 0.66
D4	**64.91** $\pm$ 0.94	52.16 $\pm$ 0.38	55.25 $\pm$ 0.90	55.40 $\pm$ 0.51	54.86 $\pm$ 0.43

Table 4: Datasets used in classification with pairwise constraints (We give the approximate size of constraints).

Data	Dimension	Constriants
w8a	300	$\simeq$8000
a9a	123	$\simeq$40000
gen	50	$\simeq$60000
svm	22	$\simeq$40000

6.2. Applications

In this subsection, we give the introduction of the applications used in our experiments.

Classification with Pairwise Constraints We evaluate the performance of all the methods on the binary classification with pairwise constraints learning problem. Given a set of training samples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$. In this task, we learn a linear model $h(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{x}^T\boldsymbol{w}$ to classify the dataset and ensure the any positive sample $\boldsymbol{x}_i^+ \in \mathcal{D}^+ := \{(\boldsymbol{x}_i, +1)\}_{i=1}^{n_p}$ has larger function value than the negative sample $\boldsymbol{x}_j^- \in \mathcal{D}^- := \{(\boldsymbol{x}_j, +1)\}_{i=1}^{n_n}$, where n_p and n_n denotes the number of positive samples and negative samples, respectively. Then, we can formulate this problem as follows,

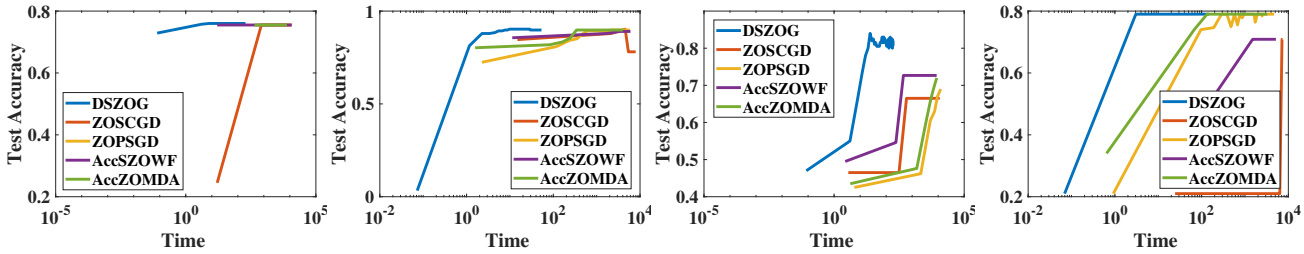
$$\begin{aligned} &\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^n \ell(h(\boldsymbol{x}_i, \boldsymbol{w}), y_i), \\ &s.t.\ h(\boldsymbol{x}_i^+, \boldsymbol{w}) - h(\boldsymbol{x}_j^-, \boldsymbol{w}) \geq 0, \\ &\quad \forall \boldsymbol{x}_i^+ \in \mathcal{D}^+\ \boldsymbol{x}_j^- \in \mathcal{D}^-, \end{aligned} \tag{23}$$

where $\ell(u, v) = c^2(1 - \exp(-\frac{(v-u)^2}{c^2}))$ is viewed as a black-box function. We summarized the datasets used in this application in Table 4. We randomly sample 1000 data samples from the original datasets, and then divide all the datasets into 3 parts, i.e., 50% for training, 30% for testing and 20% for validation. We fix the batch size of data sample at 128 for all the methods and $|\mathcal{M}_2| = |\mathcal{M}_3| = 128$. The learning rates of all the methods are chosen from $\{0.01, 0.001, 0.0001\}$. In our methods, the penalty parameter β is chosen from $\{0.1, 1, 10\}$, a and b are chosen from $\{0.1, 0.5, 0.9\}$ on the validation sets.

Classification with Fairness Constraints. In this problem, we consider the binary classification problem with a large amount of fairness constraints (Zafar et al., 2017). Given a set of training samples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. In this task, we learn a linear model $h(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{x}^T\boldsymbol{w}$. Assume that each sample has an associate sensitive feature vector $\boldsymbol{z} \in \mathbb{R}^r$. We denote $z_{ij} \in \{0, 1\}$ as the j-th sensitive feature of i-th sample. The classifier h cannot use the protected characteristic $\boldsymbol{z}$ at decision time, as it will constitute an unfair treatment. A number of metrics have been used to determine how fair a classifier is with respect to the sensitive features. According to (Zafar et al., 2017), the fair classification problems can be formulated as follows,

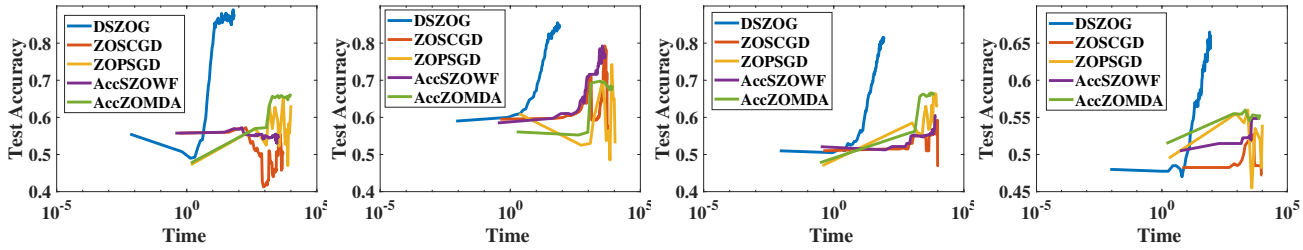
$$\begin{aligned} &\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^n \ell(h(\boldsymbol{x}_i, \boldsymbol{w}), y_i), \\ &s.t.\ \frac{1}{n}\sum_{i=1}^n (z_{ij} - \bar{z}_j) g(y_i, \boldsymbol{x}_i) \leq c, \\ &\quad \frac{1}{n}\sum_{i=1}^n (z_{ij} - \bar{z}_j) g(y_i, \boldsymbol{x}_i) \geq -c, \end{aligned} \tag{24}$$

where $j = 1, \cdots, r$, $\ell(u, v)$ denotes the loss functions, c is the covariance threshold which specifies an upper bound on the covariance between the sensitive attributes $\boldsymbol{z}$ and the signed distance $g(y, \boldsymbol{x})$. We use the hinge loss $\ell(u, v) = \max\{1 - uv, 0\}$ in this experiment and we view it as a black-box function. In addition, we use the following



(a) a9a (b) w8a (c) gen (d) svm

Figure 1: Test accuracy against training time of all the methods in classification with pairwise constraints (We stop the algorithms if the training time is more than 10000 seconds).



(a) D1 (b) D2 (c) D3 (d) D4

Figure 2: Test accuracy against training time of all the methods in classification with fairness constraints (We stop the algorithms if the training time is more than 10000 seconds).

Table 5: Datasets used in classification with fairness constraints.

Data	Dimension	Sensitive Features	Constraints
D1	100	10	40
D2	200	20	80
D3	300	20	80
D4	400	20	80

two functions to build the fairness constraints, $g(y, \boldsymbol{x}) = \begin{cases} \min\{0, \frac{1+y}{2} y h(\boldsymbol{x}, \boldsymbol{w})\} \\ \min\{0, \frac{1-y}{2} h(\boldsymbol{x}, \boldsymbol{w})\} \end{cases}$. Since the datasets with multiple sensitive features are difficult to find, we generate 4 datasets with 2000 samples in this task and summarize them in Table 5. For each dataset, we randomly choose several features to be the sensitive features, and then separate them into 3 parts, i.e., 50% for training, 30% for testing and 20% for validation. We fix the batch size of data sample at 128 for all the methods and $|\mathcal{M}_2| = |\mathcal{M}_3| = 10$. The learning rates of all the methods are chosen from $\{0.01, 0.001, 0.0001\}$. For our methods, the penalty parameter β is chosen from $\{0.1, 1, 10\}$, a and b are chosen from $\{0.1, 0.5, 0.9\}$ on the validation sets. We run all the methods 10 times on a 3990x workstation.

6.3. Results and Discussion

We present the results in Figures 1, 3 and Tables 2, 3. Note that for ZOSCGD, ZOPSGD, AccSZOFW and AccZOMDA, if the training time is larger than 10000 seconds, the algorithms are stopped. From Tables 2 and 3, we can find that our methods DSZOG has the highest test accuracy in most cases in both two applications. In addition, from Figures 1 and 3, we can find that our methods are faster than other methods. This is because all the other methods need to solve a subproblem with a large number of constraints in each iteration and the existing Python package cannot efficiently deal with such a problem. What's worse, ZOSCGD, ZOPSGD, AccSZOFW and AccZOMDA focus on solving the problem with convex constraints while the constraints in the fairness problem are nonconvex. This makes ZOSCGD, ZOPSGD, AccSZOFW and AccZOMDA cannot find the stationary point. However, by using the penalty framework, our methods can still converge to the stationary point when the constraints are nonconvex. In addition, by using a stochastic manner on the constraint, our method can efficiently deal with a large number of constraints. All these results demonstrate that our method is superior to ZOSCGD and ZOPSGD in the heavily constrained nonconvex problem.

7. Conclusion

In this paper, we propose two efficient ZO method to solve the heavily constrained nonconvex black-box problem, i.e., DSZOG. We add an additional stochastic layer into the constraint to estimate the zeroth-order gradients. In addition, momentum and adaptive step size is also used in our method. We give the convergence analysis of our proposed method. The experimental results on two applications demonstrate the superiority of our method in terms of accuracy and training time .

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant 62076138, Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX21_0999.

References

Agarwal, A., Dekel, O., and Xiao, L. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pp. 28–40. Citeseer, 2010.

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Balasubramanian, K. and Ghadimi, S. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3459–3468, 2018.

Chen, J., Zhou, D., Yi, J., and Gu, Q. A frank-wolfe framework for efficient and effective adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3486–3494, 2020.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Choromanski, K., Rowland, M., Sindhwani, V., Turner, R., and Weller, A. Structured evolution with compact architectures for scalable policy optimization. In *International Conference on Machine Learning*, pp. 970–978. PMLR, 2018.

Clarkson, K. L., Hazan, E., and Woodruff, D. P. Sublinear optimization for machine learning. *Journal of the ACM (JACM)*, 59(5):1–49, 2012.

Cotter, A., Gupta, M., and Pfeifer, J. A light touch for heavily constrained sgd. In *Conference on Learning Theory*, pp. 729–771. PMLR, 2016.

Cutkosky, A. and Mehta, H. Momentum improves normalized sgd. In *International Conference on Machine Learning*, pp. 2260–2268. PMLR, 2020.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. *Advances in Neural Information Processing Systems*, 32:15236–15245, 2019.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pp. 1646–1654, 2014.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: near-optimal non-convex optimization via stochastic path integrated differential estimator. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 687–697, 2018.

Gao, H. and Huang, H. Can stochastic zeroth-order frank-wolfe method converge faster for non-convex problems? In *International Conference on Machine Learning*, pp. 3377–3386. PMLR, 2020.

Gao, X., Jiang, B., and Zhang, S. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, 2018.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

Gu, B., Huo, Z., and Huang, H. Zeroth-order asynchronous doubly stochastic algorithm with variance reduction. *arXiv preprint arXiv:1612.01425*, 2016.

Gu, B., Liu, G., Zhang, Y., Geng, X., and Huang, H. Optimizing large-scale hyperparameters via automated learning algorithm. *arXiv preprint arXiv:2102.09026*, 2021a.

Gu, B., Wei, X., Gao, S., Xiong, Z., Deng, C., and Huang, H. Black-box reductions for zeroth-order gradient algorithms to achieve lower query complexity. *Journal of Machine Learning Research*, 22(170):1–47, 2021b.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. On stochastic moving-average estimators for non-convex optimization. *arXiv preprint arXiv:2104.14840*, 2021.

Huang, F. and Huang, H. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.

Huang, F., Gao, S., Pei, J., and Huang, H. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *arXiv preprint arXiv:2008.08170*, 2020a.

Huang, F., Tao, L., and Chen, S. Accelerated stochastic gradient-free and projection-free methods. In *International Conference on Machine Learning*, pp. 4519–4530. PMLR, 2020b.

Huang, F., Li, J., and Huang, H. Super-adam: faster and universal framework of adaptive gradients. *Advances in Neural Information Processing Systems*, 34, 2021.

Ji, K., Wang, Z., Zhou, Y., and Liang, Y. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International conference on machine learning*, pp. 3100–3109. PMLR, 2019.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

Lei, L., Ju, C., Chen, J., and Jordan, M. I. Non-convex finite-sum optimization via scsg methods. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2345–2355, 2017.

Lin, Q., Lu, Z., and Xiao, L. An accelerated proximal coordinate gradient method. *Advances in Neural Information Processing Systems*, 27:3059–3067, 2014.

Lin, Q., Ma, R., and Xu, Y. Inexact proximal-point penalty methods for constrained non-convex optimization. *arXiv preprint arXiv:1908.11518*, 2019.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Liu, M., Zhang, W., Orabona, F., and Yang, T. Adam $^+$: A stochastic method with adaptive variance reduction. *arXiv preprint arXiv:2011.11985*, 2020.

Liu, S., Chen, J., Chen, P.-Y., and Hero, A. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In *International Conference on Artificial Intelligence and Statistics*, pp. 288–297. PMLR, 2018a.

Liu, S., Kailkhura, B., Chen, P.-Y., Ting, P., Chang, S., and Amini, L. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31:3727–3737, 2018b.

Liu, S., Li, X., Chen, P.-Y., Haupt, J., and Amini, L. Zeroth-order stochastic projected gradient descent for nonconvex optimization. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1179–1183. IEEE, 2018c.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pp. 2613–2621. PMLR, 2017.

Roux, N. L., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. *arXiv preprint arXiv:1202.6258*, 2012.

Shi, W. and Gu, B. Improved penalty method via doubly stochastic gradients for bilevel hyperparameter optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9621–9629, 2021.

Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, pp. 1–67, 2021.

Wang, M., Fang, E. X., and Liu, H. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

Wang, Y., Du, S., Balakrishnan, S., and Singh, A. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1356–1365. PMLR, 2018.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. Spiderboost and momentum: Faster variance reduction algorithms. *Advances in Neural Information Processing Systems*, 32:2406–2416, 2019.

Wang, Z., Balasubramanian, K., Ma, S., and Razaviyayn, M. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *stat*, 1050:22, 2020.

Wei, X., Gu, B., and Huang, H. An accelerated variance-reduced conditional gradient sliding algorithm for first-order and zeroth-order optimization. *arXiv preprint arXiv:2109.08858*, 2021.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduction for nonconvex optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3925–3936, 2018.

A. Detailed Proofs

A.1. Proof of Proposition 5.5

Proof. Since the $(\boldsymbol{w}^*, \boldsymbol{p}^*)$ is the ϵ-stationary point of $\min_{\boldsymbol{w}} \max_{\boldsymbol{p}\in\Delta^m} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$, then we have

$$\|\nabla_{\boldsymbol{w}} f_0(\boldsymbol{w}^*) + \beta \sum_{j=1}^{m} p_j^* 2 \max\{f_j(\boldsymbol{w}^*), 0\} \nabla_{\boldsymbol{w}} f_j(\boldsymbol{w}^*)\|_2^2 \leq \epsilon^2. \tag{25}$$

Let $\alpha_j^* = 2\beta p_j^* \max\{f_j(\boldsymbol{w}^*), 0\}$ and $\epsilon \leq \epsilon_1$, we have

$$\|\nabla_{\boldsymbol{w}} f_0(\boldsymbol{w}^*) + \sum_{j=1}^{m} \alpha_j^* \nabla_{\boldsymbol{w}} f_j(\boldsymbol{w}^*)\|_2^2 \leq \epsilon_1^2. \tag{26}$$

Then the first condition in Definition 2 is satisfied.

Using $\|\nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}^*, \boldsymbol{p}^*)\|_2^2 \leq \epsilon^2$ and $0 \leq p_j^2 \leq 1$, we have

$$\sum_{j=1}^{m} (\beta \phi_j(\boldsymbol{w}^*) - \lambda p_j^*)^2 \leq \epsilon^2. \tag{27}$$

Using the inequality $\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$, we have

$$\begin{aligned} &\frac{1}{2}\beta^2 \sum_{j=1}^{m} \phi_j(\boldsymbol{w}^*)^2 \\ \leq &\sum_{j=1}^{m} (\beta \phi_j(\boldsymbol{w}^*) - \lambda p_j^*)^2 + \lambda^2 \sum_{j=1}^{m} (p_j^*)^2 \\ \leq &\epsilon^2 + m\lambda^2. \end{aligned} \tag{28}$$

Then, using $(\frac{\sum_{i=0}^{n} a_i}{n})^2 \leq \frac{\sum_{i=0}^{n} a_i^2}{n}$, we have

$$\sum_{j=1}^{m} \phi_j(\boldsymbol{w}^*) \leq \sqrt{m \sum_{j=1}^{m} \phi_j(\boldsymbol{w}^*)^2} \leq \sqrt{\frac{2m\epsilon^2 + 2m^2\lambda^2}{\beta^2}}. \tag{29}$$

Let $\sqrt{\frac{2m\epsilon^2 + 2m^2\lambda^2}{\beta^2}} \leq \epsilon_2^2$ and $\phi_j(\boldsymbol{w}) = (\max\{f_j(\boldsymbol{w}), 0\})^2$, we can obtain

$$\sum_{j=1}^{m} (\max\{f_j(\boldsymbol{w}^*), 0\})^2 \leq \epsilon_2^2. \tag{30}$$

Therefore, the second condition in Definition 2 is satisfied.

Based on the inequality $\|\langle \boldsymbol{a}, \boldsymbol{b}\rangle\|_2^2 \leq \|\boldsymbol{a}\|_2^2 \|\boldsymbol{b}\|_2^2$, we can multiply $\sum_{j=1}^{m} (\alpha_j^*)^2$ on both sides of the inequality 30, such that we have

$$\left(\sum_{j=1}^{m} \alpha_j^* \max\{f_j(\boldsymbol{w}^*), 0\}\right)^2 \leq \sum_{j=1}^{m} (\alpha_j^*)^2 \sum_{j=1}^{m} \max\{f_j(\boldsymbol{w}^*), 0\}^2 \leq \epsilon_2^2 \sum_{j=1}^{m} (\alpha_j^*)^2. \tag{31}$$

Since $\alpha_j^* \geq 0$ and $\max\{f_j(\boldsymbol{w}^*), 0\} \geq 0$

$$\sum_{j=1}^{m} (\alpha_j^* \max\{f_j(\boldsymbol{w}^*), 0\})^2 \leq \left(\sum_{j=1}^{m} \alpha_j^* \max\{f_j(\boldsymbol{w}^*), 0\}\right)^2 \leq \epsilon_2^2 \sum_{j=1}^{m} (\alpha_j^*)^2. \tag{32}$$

Using inequality 30, we have $(\alpha_j^*)^2 = 4\beta^2(p_j^*)^2(\max\{f_j(\boldsymbol{w}^*), 0\})^2 \leq 4\beta^2\epsilon_2^2$, Let $4\beta^2\epsilon_2^2 \leq \epsilon_3^2$, we have

$$\sum_{j=1}^{m}(\alpha_j^* \max\{f_j(\boldsymbol{w}^*), 0\})^2 \leq \epsilon_3^2. \tag{33}$$

If $f_j(\boldsymbol{w}^*) \leq 0$, we have $\alpha_j^* = 2\beta p_j^* \max\{f_j(\boldsymbol{w}^*), 0\} = 0$. Therefore, we have

$$\sum_{j=1}^{m}(\alpha_j^* f_j(\boldsymbol{w}^*))^2 \leq \epsilon_3^2, \tag{34}$$

which means that the third condition in Definition 2 is satisfied.

That completes the proof. □

A.2. Proof of Proposition 5.7

Proof. Assume that a point $\hat{\boldsymbol{w}}$ satisfies that $\|\nabla_{\boldsymbol{w}} g(\hat{\boldsymbol{w}})\|_2 \leq \epsilon$, the optimization problem $\max_{\boldsymbol{p}\in\Delta^m} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p})$ is strongly concave w.r.t $\boldsymbol{p}$ and $\boldsymbol{p}^*(\hat{\boldsymbol{w}})$ is uniquely defined. Solving this this strongly concave problem $\max_{\boldsymbol{p}\in\Delta^m} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p})$, we can obtain a point $\boldsymbol{p}'$ satisfying that

$$\|\nabla_{\boldsymbol{p}} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p}')\|_2 \leq \epsilon \; and \; \|\boldsymbol{p}' - \boldsymbol{p}^*(\hat{\boldsymbol{w}})\|_2 \leq \epsilon. \tag{35}$$

If $\|\nabla_{\boldsymbol{w}} g(\hat{\boldsymbol{w}})\|_2 \leq \epsilon$, we have

$$\begin{aligned}
&\|\nabla_{\boldsymbol{w}} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p}')\|_2 \\
\leq &\|\nabla_{\boldsymbol{w}} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p}') - \nabla_{\boldsymbol{w}} g(\hat{\boldsymbol{w}})\|_2 + \|\nabla_{\boldsymbol{w}} g(\hat{\boldsymbol{w}})\|_2 \\
= &\|\nabla_{\boldsymbol{w}} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p}') - \nabla_{\boldsymbol{w}} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p}^*(\hat{\boldsymbol{w}}))\|_2 + \epsilon \\
\leq &L\|\boldsymbol{p}' - \boldsymbol{p}^*(\hat{\boldsymbol{w}})\|_2 + \epsilon \\
= &\mathcal{O}(\epsilon).
\end{aligned} \tag{36}$$

□

A.3. Proof of Lemma 5.11

Proof. Under Assumptions 5.1 and 5.9, we have

$$
\begin{aligned}
&g(\boldsymbol{w}_{t+1}) \\
\leq& g(\boldsymbol{w}_t) + \nabla g(\boldsymbol{w}_t)^T(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) + \frac{L}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2^2 \\
=& g(\boldsymbol{w}_t) - \eta_w \nabla g(\boldsymbol{w}_t)^T \frac{\boldsymbol{z}_{\boldsymbol{w}}^t}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2} + c} + \frac{L}{2}\frac{\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2}{\|\sqrt{\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2} + c\|_2^2} \\
\leq& g(\boldsymbol{w}_t) - \eta_w c_{1,l} \nabla g(\boldsymbol{w}_t)^T \boldsymbol{z}_{\boldsymbol{w}}^t + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
=& g(\boldsymbol{w}_t) + \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
=& g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
&+ \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t) - \nabla g_\mu(\boldsymbol{w}_t) + \nabla g_\mu(\boldsymbol{w}_t) - \nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) - \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) \\
&+ \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
\leq& g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
&+ \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) - \nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t))\|_2^2 + \eta_w c_{1,l}\|\nabla\nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) - \nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t)\|_2^2 \\
&+ \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) - \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)\|_2^2 + \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
\leq& g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \eta_w c_{1,l}\frac{\mu^2 L^2 (d+3)^3}{4} + \eta_w c_{1,l} L^2\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 \\
&+ \eta_w c_{1,l}\frac{\mu^2 L^2 (d+3)^3}{4} + \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
\leq& g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \eta_w c_{1,l}\frac{\mu^2 L^2 (d+3)^3}{2} + \eta_w c_{1,l} L^2\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 \\
&+ \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
\leq& g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{4}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \eta_w c_{1,l}\frac{\mu^2 L^2 (d+3)^3}{2} + \eta_w c_{1,l} L^2\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 \\
&+ \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2.
\end{aligned}
\tag{37}
$$

The last inequality is due to $\eta_w L \leq \frac{c_{1,l}}{2c_{1,u}^2}$. □

A.4. Proof of Lemma 5.12

Proof. According to the update rule of $\boldsymbol{p}$, we have

$$
\begin{aligned}
&\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 \\
\leq& \|(1-a)\boldsymbol{p}_t + a\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 \\
=& \|\boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 + a^2\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 + 2a\langle \boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t), \boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\rangle.
\end{aligned}
\tag{38}
$$

Rearrange the above inequality, we have

$$
\langle \boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t), \boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\rangle \geq \frac{1}{2a}\left(\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 - \|\boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 - a^2\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2\right). \tag{39}
$$

Due to the Assumption 5.1, we have

$$
\mathcal{L}(\boldsymbol{w}_t, \hat{\boldsymbol{p}}_{t+1}) \geq \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T(\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t) - \frac{L}{2}\|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2. \tag{40}
$$

In addition, according to the strongly concave, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}) \\
\leq&\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T(\boldsymbol{p} - \boldsymbol{p}_t) - \frac{\tau}{2}\|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2 \\
=&\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T(\boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1} + \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t) - \frac{\tau}{2}\|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2 \\
=&\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T(\boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}) + \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T(\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t) - \frac{\tau}{2}\|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2 \\
=&\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + (\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t)^T(\boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}) + \langle \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}\rangle + \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T(\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t) - \frac{\tau}{2}\|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2. \quad (41)
\end{aligned}$$

Then, using the above inequalities, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}) \\
\leq&\mathcal{L}(\boldsymbol{w}_t, \hat{\boldsymbol{p}}_{t+1}) + (\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t)^T(\boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}) + \langle \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}\rangle - \frac{\tau}{2}\|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2 + \frac{L}{2}\|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2. \quad (42)
\end{aligned}$$

Due to the update rule of $\hat{\boldsymbol{p}}$, we have

$$\langle \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t - \eta_{\boldsymbol{p}}\frac{\boldsymbol{z}_{\boldsymbol{p}}^t}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{p}}^t\|_2} + c}, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}\rangle \geq 0, \quad \forall \boldsymbol{p} \in \Delta^m. \quad (43)$$

Then, we have

$$\begin{aligned}
&\eta_{\boldsymbol{p}}c_{2,l}\langle \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}\rangle \\
\leq&\langle \eta_{\boldsymbol{p}}\frac{\boldsymbol{z}_{\boldsymbol{p}}^t}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{p}}^t\|_2} + c}, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}\rangle \\
\leq&\langle \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}\rangle \\
=&\langle \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t, \boldsymbol{p} - \boldsymbol{p}_t + \boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\rangle \\
=&-\|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2 + \langle \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t, \boldsymbol{p} - \boldsymbol{p}_t\rangle. \quad (44)
\end{aligned}$$

In addition, we have

$$\begin{aligned}
&\langle \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p}^*(\boldsymbol{w}_t) - \hat{\boldsymbol{p}}_{t+1}\rangle \\
=&\langle \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\rangle + \langle \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\rangle \\
\leq&\frac{1}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + \frac{\tau}{4}\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{1}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + \frac{\tau}{4}\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 \\
\leq&\frac{2}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + \frac{\tau}{4}\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{\tau}{4}\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2. \quad (45)
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}) \\
\leq&\mathcal{L}(\boldsymbol{w}_t, \hat{\boldsymbol{p}}_{t+1}) - \frac{1}{\eta_p c_{2,l}}\|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2 + \frac{1}{\eta_p c_{2,l}}\langle \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t, \boldsymbol{p} - \boldsymbol{p}_t\rangle - \frac{\tau}{2}\|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2 + \frac{L}{2}\|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2 \\
&+ \frac{2}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + \frac{\tau}{4}\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{\tau}{4}\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2. \quad (46)
\end{aligned}$$

Let $\boldsymbol{p} = \boldsymbol{p}^*(\boldsymbol{w}_t)$, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_t, \hat{\boldsymbol{p}}_{t+1}) \\
\leq &\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) \\
\leq &\mathcal{L}(\boldsymbol{w}_t, \hat{\boldsymbol{p}}_{t+1}) - \frac{1}{\eta_p c_{2,l}}\|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2 - \frac{\tau}{2}\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{L}{2}\|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2 \\
&+ \frac{2}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + \frac{\tau}{4}\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{\tau}{4}\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 \\
&- \frac{1}{2a\eta_p c_{2,l}}\left(\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 - \|\boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 - a^2\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2\right).
\end{aligned} \tag{47}$$

Rearrange the inequality, we have

$$\begin{aligned}
&\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 \\
\leq &-2a\eta_p c_{2,l}\left(\frac{1}{\eta_p c_{2,l}} - \frac{L}{2} - \frac{\tau}{4} - \frac{1}{2b\eta_p c_{2,l}}\right)\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 + \frac{4a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{2})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 \\
\leq &-2a\eta_p c_{2,l}\left(\frac{1}{2\eta_p c_{2,l}} - \frac{3L}{4}\right)\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 + \frac{4a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{2})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2.
\end{aligned} \tag{48}$$

where we use $a \leq 1$, $\tau \leq L$ and $\eta_p \leq \frac{1}{3c_{2.l}L}$.

Then, we have

$$\begin{aligned}
&\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2 \\
= &\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t) + \boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2 \\
\leq &(1 + \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 + (1 + \frac{4}{\tau a\eta_p c_{2,l}})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2 \\
\leq &-2a\eta_p c_{2,l}(1 + \frac{\tau a\eta_p c_{2,l}}{4})\left(\frac{1}{2\eta_p c_{2,l}} - \frac{3L}{4}\right)\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 + \frac{4a\eta_p c_{2,l}}{\tau}(1 + \frac{\tau a\eta_p c_{2,l}}{4})\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{2})(1 + \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + (1 + \frac{4}{\tau a\eta_p c_{2,l}})L_g^2\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2^2 \\
\leq &-2a\eta_p c_{2,l}(1 + \frac{\tau a\eta_p c_{2,l}}{4})\left(\frac{1}{2\eta_p c_{2,l}} - \frac{3L}{4}\right)\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 + \frac{8a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{8L_g^2}{\tau a\eta_p c_{2,l}}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2^2 \\
\leq &-\frac{2\eta_p c_{2,l}}{a}(1 + \frac{\tau a\eta_p c_{2,l}}{4})\left(\frac{1}{2\eta_p c_{2,l}} - \frac{3L}{4}\right)\|\boldsymbol{p}_t - \boldsymbol{p}_{t+1}\|_2^2 + \frac{8a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{8L_g^2}{\tau a\eta_p c_{2,l}}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2^2 \\
\leq &-\frac{1}{4a}\|\boldsymbol{p}_t - \boldsymbol{p}_{t+1}\|_2^2 + \frac{8a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{8L_g^2}{\tau a\eta_p c_{2,l}}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2^2.
\end{aligned} \tag{49}$$

□

A.5. Proof of Lemma 5.13

Proof. According to the update rule of $z_{\boldsymbol{p}}$, we have

$$\boldsymbol{z}_{\boldsymbol{p}}^{t+1} - \boldsymbol{z}_{\boldsymbol{p}}^{t} = -b\boldsymbol{z}_{\boldsymbol{p}}^{t} + bH^{t+1}. \tag{50}$$

Then, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \boldsymbol{z}_{\boldsymbol{p}}^{t+1}\|_2^2] \\
=&\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \boldsymbol{z}_{\boldsymbol{p}}^{t} - (\boldsymbol{z}_{\boldsymbol{p}}^{t+1} - \boldsymbol{z}_{\boldsymbol{p}}^{t})\|_2^2] \\
=&\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \boldsymbol{z}_{\boldsymbol{p}}^{t} + b\boldsymbol{z}_{\boldsymbol{p}}^{t} - bH^{t+1}\|_2^2] \\
=&\mathbb{E}[\|(1-b)\left(\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t}) - \boldsymbol{z}_{\boldsymbol{p}}^{t}\right) + (1-b)\left(\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t})\right) + b\left(\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - H^{t+1}\right)\|_2^2] \\
=&(1-b)^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t}) - \boldsymbol{z}_{\boldsymbol{p}}^{t}\|_2^2] + (1-b)^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t})\|_2^2] \\
&+ b^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - H^{t+1}\|_2^2] + (1-b)^2\langle\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t}) - \boldsymbol{z}_{\boldsymbol{p}}^{t}, \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t})\rangle \\
=&(1-b)^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t}) - \boldsymbol{z}_{\boldsymbol{p}}^{t}\|_2^2] + (1-b)^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t})\|_2^2] \\
&+ b^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - H^{t+1}\|_2^2] + (1-b)^2\mathbb{E}[\langle\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t}) - \boldsymbol{z}_{\boldsymbol{p}}^{t}, \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t})\rangle] \\
\leq&(1-b)^2(1+b)\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t}) - \boldsymbol{z}_{\boldsymbol{p}}^{t}\|_2^2] + (1-b)^2(1+\frac{1}{b})\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t})\|_2^2 \\
&+ b^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - H^{t+1}\|_2^2] \\
\leq&(1-b)\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t}) - \boldsymbol{z}_{\boldsymbol{p}}^{t}\|_2^2] + \frac{1}{b}L^2\mathbb{E}[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2^2 + \|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\|_2^2] + b^2\sigma_2^2.
\end{aligned} \tag{51}$$

Similarly, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \boldsymbol{z}_{\boldsymbol{w}}^{t+1}\|_2^2] \\
\leq&(1-b)\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t}) - \boldsymbol{z}_{\boldsymbol{w}}^{t}\|_2^2] + \frac{1}{b}L^2\mathbb{E}[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2^2 + \|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\|_2^2] + b^2\sigma_1^2.
\end{aligned} \tag{52}$$

□

A.6. Proof of Theorem 5.14

Proof. Summing up the inequality in Lemma 5.13, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \boldsymbol{z}_{\boldsymbol{w}}^{t+1}\|_2^2] \\
\leq&(1-b)\sum_{t=1}^{T}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{t},\boldsymbol{p}_{t}) - \boldsymbol{z}_{\boldsymbol{w}}^{t}\|_2^2] + \frac{1}{b}L^2\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2^2 + \|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\|_2^2] + b^2\sigma_1^2T.
\end{aligned} \tag{53}$$

Then, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \boldsymbol{z}_{\boldsymbol{w}}^{t+1}\|_2^2] \\
\leq&\frac{1}{b}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{1},\boldsymbol{p}_{1}) - \boldsymbol{z}_{\boldsymbol{w}}^{1}\|_2^2] + \frac{1}{b^2}L^2\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2^2 + \|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\|_2^2] + b\sigma_1^2T.
\end{aligned} \tag{54}$$

Similarly, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1}) - \boldsymbol{z}_{\boldsymbol{p}}^{t+1}\|_2^2] \\
\leq&\frac{1}{b}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{1},\boldsymbol{p}_{1}) - \boldsymbol{z}_{\boldsymbol{w}}^{1}\|_2^2] + \frac{1}{b^2}L^2\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2^2 + \|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\|_2^2] + b\sigma_1^2T.
\end{aligned} \tag{55}$$

$$
\begin{aligned}
&\sum_{t=1}^{T}\|\boldsymbol{p}_{t+1}-\boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2\\
\leq&\frac{4}{\tau a\eta_p c_{2,l}}(\|\boldsymbol{p}^*(\boldsymbol{w}_1)-\boldsymbol{p}_1\|_2^2-\frac{1}{4a}\sum_{t=1}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_2^2+\frac{8a\eta_p c_{2,l}}{\tau}\sum_{t=1}^{T}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2\\
&+\frac{8L_g^2}{\tau a\eta_p c_{2,l}}\sum_{t=1}^{T}\|\boldsymbol{w}_t-\boldsymbol{w}_{t+1}\|_2^2)\\
\leq&\frac{4}{\tau a\eta_p c_{2,l}}\|\boldsymbol{p}^*(\boldsymbol{w}_1)-\boldsymbol{p}_1\|_2^2-\frac{1}{\tau a^2\eta_p c_{2,l}}\sum_{t=1}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_2^2+\frac{32}{\tau^2}\sum_{t=1}^{T}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2+\frac{32L_g^2\eta_w^2c_{1,u}^2}{\tau^2a^2\eta_p^2c_{2,l}^2}\sum_{t=1}^{T}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2.
\end{aligned}
\tag{56}
$$

Thus, we have

$$
\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T}\|\boldsymbol{p}_{t+1}-\boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2]\\
\leq&\frac{4}{\tau a\eta_p c_{2,l}}\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_1)-\boldsymbol{p}_1\|_2^2]+\mathbb{E}[\frac{32L_g^2\eta_w^2c_{1,u}^2}{\tau^2a^2\eta_p^2c_{2,l}^2}\sum_{t=1}^{T}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2-\frac{1}{\tau a^2\eta_p c_{2,l}}\sum_{t=1}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_2^2]\\
&+\frac{32}{\tau^2 b}\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)-\boldsymbol{z}_{\boldsymbol{p}}^1\|_2^2]+\mathbb{E}\left[\sum_{t=1}^{T}\frac{32}{\tau^2}\left(b\sigma_2^2+\frac{L^2(\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2^2+\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|_2^2)}{b^2}\right)\right]\\
\leq&\frac{4}{\tau a\eta_p c_{2,l}}\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_1)-\boldsymbol{p}_1\|_2^2]+\mathbb{E}\left[\left(\frac{32L_g^2\eta_w^2c_{1,u}^2}{\tau^2a^2\eta_p^2c_{2,l}^2}+\frac{32L^2\eta_w^2c_{1,u}^2}{\tau^2b^2}\right)\sum_{t=1}^{T}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2\right]\\
&+\frac{32}{\tau^2 b}\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)-\boldsymbol{z}_{\boldsymbol{p}}^1\|_2^2]+\mathbb{E}\left[\sum_{t=1}^{T}\frac{32}{\tau^2}b\sigma_2^2\right]\\
&+\mathbb{E}\left[\left(\frac{32L^2}{\tau^2b^2}-\frac{1}{\tau a^2\eta_p c_{2,l}}\right)\sum_{t=1}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_2^2]\right].
\end{aligned}
\tag{57}
$$

In addition, we have

$$
\begin{aligned}
&\frac{1}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2\\
\leq&\frac{g(\boldsymbol{w}_t)-g(\boldsymbol{w}_{t+1})}{\eta_w c_{1,l}}-\frac{1}{4}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2+\frac{\mu^2L^2(d+3)^3}{2}+L^2\|\boldsymbol{p}^*(\boldsymbol{w}_t)-\boldsymbol{p}_t\|_2^2+\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2.
\end{aligned}
\tag{58}
$$

Summing up from $t = 1, \cdots, T$ and taking expectation, we have

$$\begin{aligned}
&\mathbb{E}[\frac{1}{2}\sum_{t=1}^{T}\|\nabla g(\boldsymbol{w}_t)\|_2^2] \\
\leq&\frac{g(\boldsymbol{w}_1)-g(\boldsymbol{w}_{T+1})}{\eta_w c_{1,l}} - \frac{1}{4}\mathbb{E}[\sum_{t=1}^{T}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2] + \frac{\mu^2 T L^2 (d+3)^3}{2} \\
&+ \frac{4L^2}{\tau a \eta_p c_{2,l}}\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_1)-\boldsymbol{p}_1\|_2^2] + \mathbb{E}\left[\left(\frac{L^2\eta_w^2 c_{1,u}^2}{b^2} + \frac{32L_g^2 L^2 \eta_w^2 c_{1,u}^2}{\tau^2 a^2 \eta_p^2 c_{2,l}^2} + \frac{32L^4\eta_w^2 c_{1,u}^2}{\tau^2 b^2}\right)\sum_{t=1}^{T}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2\right] \\
&+ \frac{32L^2}{\tau^2 b}\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)-\boldsymbol{z}_{\boldsymbol{p}}^1\|_2^2] + \mathbb{E}\left[\sum_{t=1}^{T}\frac{32L^2}{\tau^2}b\sigma_2^2\right] \\
&+ \mathbb{E}\left[\left(\frac{L^2}{b^2} + \frac{32L^4}{\tau^2 b^2} - \frac{L^2}{\tau a^2 \eta_p c_{2,l}}\right)\sum_{t=1}^{T}\|\boldsymbol{p}_t - \boldsymbol{p}_{t+1}\|_2^2\right] \\
&+ \frac{1}{b}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)-\boldsymbol{z}_{\boldsymbol{w}}^1\|_2^2] + b\sigma_1^2 T. \qquad (59)
\end{aligned}$$

Let $\boldsymbol{p}^*(w_1) = \boldsymbol{p}_1$, $\boldsymbol{z}_{\boldsymbol{p}}^1 = H(\boldsymbol{w}_t, \boldsymbol{p}_t, f_{\mathcal{M}_3})$, $\boldsymbol{z}_{\boldsymbol{w}}^1 = G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})$, $\eta_p \leq \min\{\frac{b^2}{\tau a^2 c_{2,l}}, \frac{\tau b^2}{32L^2 a^2 c_{2,l}}\}$, $\eta_w^2 \leq \min\{\frac{b^2}{4c_{1,u}^2 L^2}, \frac{\tau^2 a^2 \eta_p^2 c_{2,l}^2}{128 L_g^2 L^2 c_{1,u}}, \frac{\tau^2 b^2}{128 L^4 c_{1,u}^2}\}$, we have

$$\frac{1}{T}\mathbb{E}[\sum_{t=1}^{T}\|\nabla g(\boldsymbol{w}_t)\|_2^2] \leq \frac{2(g(\boldsymbol{w}_1)-g(\boldsymbol{w}_{T+1}))}{T\eta_w c_{1,l}} + \frac{64L^2}{\tau^2 bT}\sigma_2^2 + \frac{2\sigma_1^2}{bT} + \mu^2 L^2 (d+3)^3 + \frac{64L^2}{\tau^2}b\sigma_2^2 + 2b\sigma_1^2. \qquad (60)$$

Bound the left term by ϵ^2, we have $\mu \leq \frac{\epsilon}{L(d+3)^{3/2}}$, $b \leq \min\{\frac{\epsilon^2}{2\sigma_1^2}, \frac{\tau^2\epsilon^2}{64\sigma_2^2 L^2}\}$ and $T \geq \max\{\frac{2(g(\boldsymbol{w}_1)-g(\boldsymbol{w}_T))}{\epsilon^2 \eta_w c_{1,l}}, \frac{2\sigma_1^2}{\epsilon^2 b}, \frac{64\sigma_2^2 L^2}{\epsilon^2 \tau^2 b}\}$.

□

B. Additional Experiments

B.1. Impact of the Hyper-parameters

In this section, we discuss the impact of the learning rate $\eta_{\boldsymbol{p}}$ and $\eta_{\boldsymbol{w}}$. First, let $\eta_{\boldsymbol{p}} = 0.001$. We evaluate the performance of our method in two applications with $\eta_{\boldsymbol{w}}$ chosen from $\{0.01, 0.001, 0.0001\}$. Then, let $\eta_{\boldsymbol{w}} = 0.01$. We evaluate the performance in two applications with $\eta_{\boldsymbol{p}}$ chosen from $\{0.01, 0.001, 0.0001\}$. All the experiments are presented in Tables 6, 8, 7 and 9. We can find that it is important to choose a proper $\eta_{\boldsymbol{w}}$. In addition, our method is not sensitive to $\eta_{\boldsymbol{p}}$. We also give the results of our method with different β a and b in Table 11 and Table 12. We can find that our method is not sensitive to β, a and b.

B.2. Performance with Nonlinear Model

We use the kernel method $k(x, x') = \exp(-\gamma\|x - x'\|_2^2)$ to conduct a nonlinear model. The hyperparameter of all the methods are set according to Section 6. The results of fairness are presented in Table 10 and Figure 3. We can find that our method is still superior than other methods.

Table 6: Test accuracy (%) of DSZOG in classification with fairness constraints when using different $\eta_{\boldsymbol{w}}$.

Data	0.01	0.001	0.0001
D1	87.33 ± 0.38	87.21 ± 0.34	80.68 ± 0.32
D2	84.75 ± 0.25	83.01 ± 0.24	67.34 ± 0.78
D3	83.58 ± 0.14	83.78 ± 0.45	77.22 ± 0.32
D4	64.91 ± 0.94	63.27 ± 0.45	55.13 ± 0.46

Table 7: Test accuracy (%) of DSZOG in classification with pairwise constraints when using different $\eta_{\boldsymbol{w}}$.

Data	0.01	0.001	0.0001
a9a	75.90 ± 0.26	74.45 ± 0.67	65.33 ± 0.43
w8a	89.94 ± 0.28	88.28 ± 0.80	56.88 ± 0.22
gen	82.33 ± 0.76	81.78 ± 0.45	57.34 ± 0.22
svm	79.56 ± 0.49	79.56 ± 0.45	56.23 ± 0.75

Table 8: Test accuracy (%) of DSZOG in classification with fairness constraints when using different $\eta_{\boldsymbol{p}}$.

Data	0.01	0.001	0.0001
D1	87.33 ± 0.38	86.12 ± 0.23	85.13 ± 0.22
D2	84.75 ± 0.25	83.22 ± 0.14	83.05 ± 0.03
D3	83.58 ± 0.14	83.29 ± 0.31	81.18 ± 0.22
D4	64.91 ± 0.94	62.12 ± 0.31	62.23 ± 0.22

Table 9: Test accuracy (%) of DSZOG in classification with pairwise constraints when using different $\eta_{\boldsymbol{p}}$.

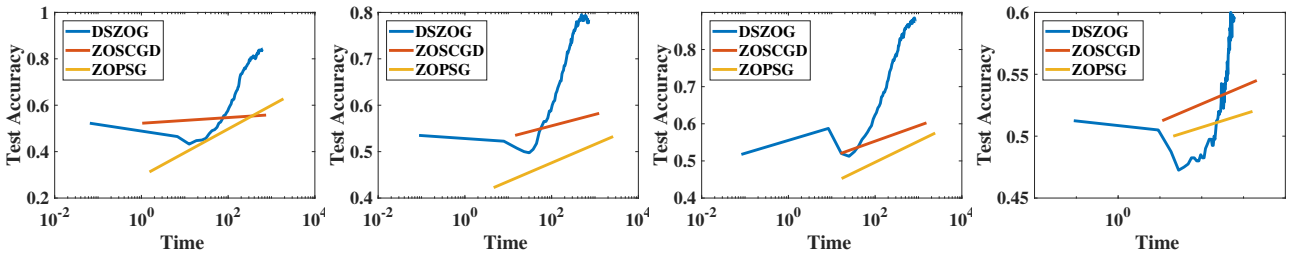
Data	0.01	0.001	0.0001
a9a	75.90 ± 0.26	73.33 ± 0.31	73.64 ± 0.22
w8a	89.94 ± 0.28	89.23 ± 0.17	88.82 ± 0.56
gen	82.33 ± 0.76	82.11 ± 0.64	81.28 ± 0.34
svm	79.56 ± 0.49	78.56 ± 0.44	78.45 ± 0.76

Table 10: Test accuracy (%) in fairness with kernel method.

Data	DSZOG	ZOSCGD	ZOPSGD
D1	**84.25**	55.75	62.75
D2	**78.50**	58.25	53.25
D3	**87.50**	60.25	57.50
D4	**59.52**	54.50	52.25

Table 11: Test accuracy of DSZOG with different β fairness ($\eta_w = 0.001$, $\eta_p = 0.1$, $a = 0.9$, $b = 0.9$).

Data	0.1	1	10
D1	84.25	84.73	83.12
D2	78.50	77.98	78.33
D3	87.50	87.67	87.85
D4	59.52	58.61	58.98



(a) D1 (b) D2 (c) D3 (d) D4

Figure 3: Performance of our method in fairness with kernel method.

Table 12: Test accuracy (%) of DSZOG with different a and b fairness constraints ($\eta_w = 0.001$, $\eta_p = 0.1$, $\beta = 0.1$).

	$a = 0.9$			$b = 0.9$		
	$b = 0.1$	$b = 0.5$	$b = 0.9$	$a = 0.1$	$a = 0.5$	$b = 0.9$
D1	83.93	84.12	84.25	82.12	83.21	84.25
D2	78.33	78.27	78.50	77.29	78.45	78.50
D3	85.34	87.410	87.50	87.23	87.14	87.50
D4	58.82	57.23	59.52	58.78	59.48	59.52