ICML International Conference On Machine Learning

Thirty-ninth International Conference on Machine Learning

Gradient-Free Method for Heavily Constrained Nonconvex Optimization

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

[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>.

Outline

- Problem setting
- Related works
- Proposed method and algorithm
- Theoretical results
- Experiments

Problem setting

- We consider the following problem

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

  - $f_0(\cdot)$ is a non-convex and white/black-box function
  - $f_j(\cdot)$ is non-convex/convex and white/black-box function

- Examples
  1. Classification with pairwise constraints.
  2. Tuning the average performance of the policy under multiple scenarios and ensuring the performance of each scenario.
  3. Optimizing the control policy under performance and safety constraints.
  4. ⋯

Related works

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

Proposed method

- Reformulate the problem as the following minimax problem over a probability distribution,

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

where $\beta > 0$, $\lambda > 0$, $\varphi(\boldsymbol{w},\boldsymbol{p}) = \sum_{j=1}^{m} p_j \phi_j(\boldsymbol{w})$, $\phi_j(\boldsymbol{w}) = \left(\max\{f_j(\boldsymbol{w}), 0\}\right)^2$, $\Delta^m := \{\boldsymbol{p} | \sum_{j=1}^{d} p_j = 1,$ $0 \le p_j \le 1\}$.

- Alternately update $\boldsymbol{w}$ and $\boldsymbol{p}$ with stochastic zeroth-order method.

Proposed method

- *Sample ℓ_i uniformly,* and *f_j according to $\boldsymbol{p}$* , and calculate their stochastic zeroth-order gradient w.r.t $\boldsymbol{w}$,

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

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

where $\mu > 0$ and $u \sim N(0, 1_d)$.

- Obtain the stochastic zeroth-order gradient of $\mathcal{L}$ w.r.t $\boldsymbol{w}$,

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

Proposed method

- *Sample f_j uniformly,* and calculate the stochastic gradient w.r.t $\boldsymbol{p}$,

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

- Sample a *batch* of l_i, f_j, and u_k to reduce the variance

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

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

where $M_1 \subseteq [n]$, $M_2 \subseteq [m]$, $M_3 \subseteq [m]$ and $q > 0$.

Proposed method

- Using *momentum methods* and *adaptive step size* to

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

$$z_{\boldsymbol{p}}^{t+1} = (1-b) z_{\boldsymbol{p}}^{t} + b H(\boldsymbol{w}_{t+1}, \boldsymbol{p}_{t+1}, f_{\mathcal{M}_3}), \tag{12}$$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_{\boldsymbol{w}} \frac{z_{\boldsymbol{w}}^t}{\sqrt{\|z_{\boldsymbol{w}}^t\|_2 + c}}, \tag{13}$$

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

where $p_{t+1} = (1-a)p_t + a\hat{p}_{t+1}$, and $P_{\Delta^m}(\cdot)$ denotes the projection operator.

Proposed method

- *Algorithm*

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

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

Output: $\boldsymbol{w}_T$.

Initialize $\boldsymbol{w}_1$.

Initialize $\boldsymbol{p}_1 = \boldsymbol{p}^*(\boldsymbol{w}_1)$ by solving the strongly concave problem.

Initialize $\boldsymbol{z}_{\boldsymbol{w}}^1 = G_\mu^{\mathcal{L}}(\boldsymbol{w}_1, \boldsymbol{p}_1, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})$ and $\boldsymbol{z}_{\boldsymbol{p}}^1 = H(\boldsymbol{w}_1, \boldsymbol{p}_1, f_{\mathcal{M}_3})$.

for $t = 1, \cdots, T$ **do**

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_{\boldsymbol{w}} \frac{\boldsymbol{z}_{\boldsymbol{w}}^t}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2 + c}}.$$

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

$\boldsymbol{p}_{t+1} = (1-a)\boldsymbol{p}_t + a\hat{\boldsymbol{p}}_{t+1}$.

Randomly sample $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_q \sim \mathcal{N}(0, \mathbf{1}_d)$.

Randomly sample a index set $\mathcal{M}_1 \subseteq [n]$ of ℓ_i.

Sample a constraint index set $\mathcal{M}_2 \sim \boldsymbol{p}_{t+1} \subseteq [m]$.

Randomly sample a constraint index set $\mathcal{M}_3$.

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

Calculate $H(\boldsymbol{w}_{t+1}, \boldsymbol{p}_{t+1}, f_{\mathcal{M}_3}) = \frac{\beta m}{|\mathcal{M}_3|} \sum_{j \in \mathcal{M}_3} e_j \phi_j(\boldsymbol{w}_{t+1}) - \lambda \boldsymbol{p}_{t+1}$.

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

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

end for

Proposed method

- Convergence analysis

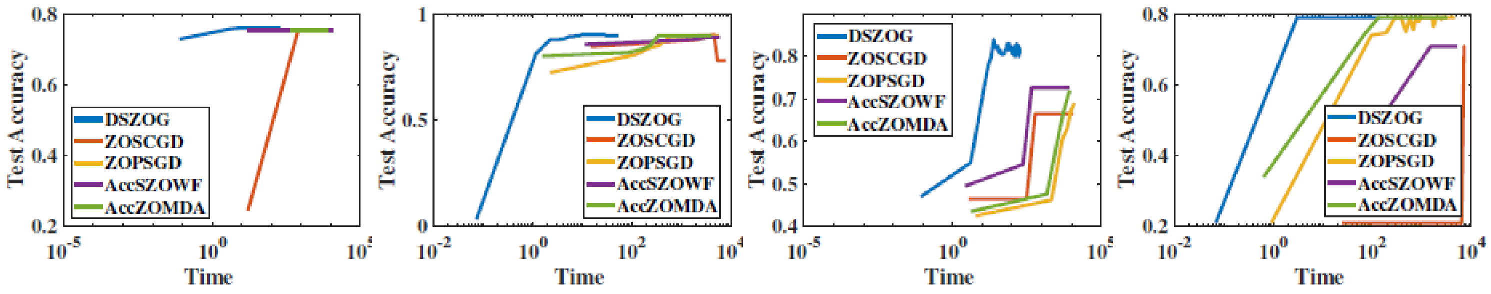
$$\min_{\boldsymbol{w}} \left\{ g(\boldsymbol{w}) := \max_{\boldsymbol{p}\in\Delta^m} \mathcal{L}(\boldsymbol{w},\boldsymbol{p}) = \mathcal{L}(\boldsymbol{w},\boldsymbol{p}^*(\boldsymbol{w})) \right\}, \quad (21)$$

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

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

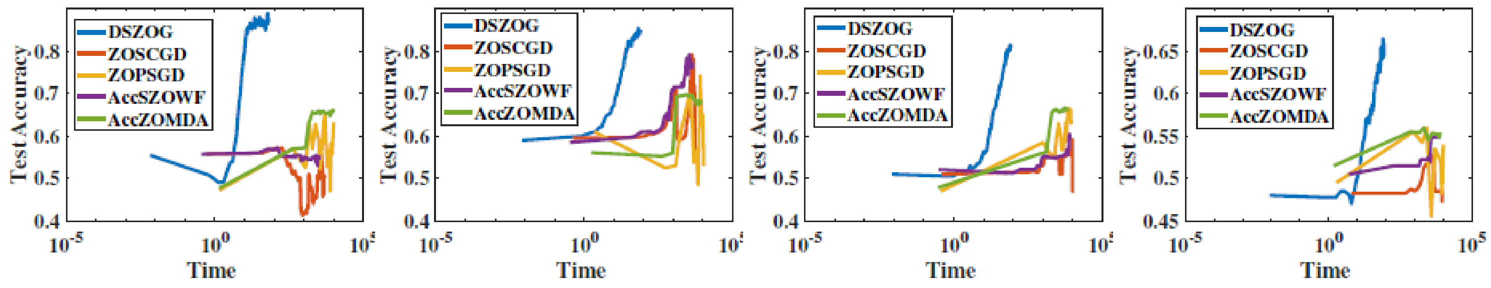
Proposed method

- Experimental results



(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).

Thank you!