Non-convex Constrained Optimization with Many Stochastic Constraints

Tianbao Yang
CSE@TAMU

Outline

- **Problem Formulation**
- **Penalty methods**
- **Applications**

Problem Formulation

Functional Inequality Constrained Optimization

$$\min_{x} \quad f_0(x)$$
$$s.t. \quad g_i(x) \leq 0, i = 1, \ldots, m$$

Non-convex **Non-smooth**

$$f_0(x) = \mathbb{E}_{\zeta}[f_0(x, \zeta)]$$

Non-convex **Non-smooth**

$$g_i(x) = \mathbb{E}_{\xi_i}[g_i(x, \xi_i)]$$

m **is large**

KKT solutions

$$\exists \lambda \geq 0$$

$$0 \in \partial f_0(x) + \sum_{k=1}^{m} \lambda_k \partial g_k(x)$$

$$g_k(x) \leq 0, \forall k$$

$$\lambda_k g_k(x) = 0, \forall k$$

Nearly epsilon-KKT solution

$$\|x - \bar{x}\|_2 \leq O(\epsilon)$$

$$\text{dist}(0, \partial f_0(\bar{x}) + \sum_{k=1}^{m} \lambda_k \partial g_k(\bar{x})) \leq \epsilon$$

$$[g_k(\bar{x})]_+ \leq \epsilon, \forall k$$

$$|\lambda_k g_k(\bar{x})| \leq \epsilon, \forall k$$

Related works

- Proximal-point based methods
  - Ma et al. (2020) and Boob et al. (2023)
- Switching gradient methods
  - Huang & Lin (2023)
- Augmented Lagrangian Method
  - Li et al. 2022
- Penalty-based methods
  - Alacaoglu & Wright (2024), Liu & Xu (2025)

Summary of Existing Results

Reference	Loop	Constraints	Smoothness	Convexity	Complexity
(Alacaoglu & Wright, 2024)	Single Loop	$g(\cdot) = 0$	f_0, g	NC (f_0, g)	$\tilde{\mathcal{O}}(\epsilon^{-5})$
(Li et al., 2024b)	Double Loop	$g(\cdot) = 0$	f_0, g	NC (f_0, g)	$\tilde{\mathcal{O}}(\epsilon^{-5})$
(Ma et al., 2020)	Double Loop	$g(\cdot) \leq 0$	none	WC (f_0, g)	$\mathcal{O}(\epsilon^{-6})$
(Boob et al., 2023)	Double Loop	$g(\cdot) \leq 0$	none	WC (f_0, g)	$\mathcal{O}(\epsilon^{-6})$
(Huang & Lin, 2023)	Single Loop	$g(\cdot) \leq 0$	none	WC (f_0), C (g)	$\tilde{\mathcal{O}}(\epsilon^{-8})$
(Liu & Xu, 2025)	Single Loop	$g(\cdot) \leq 0$	none	WC (f_0, g)	$\mathcal{O}(\epsilon^{-6})$

Gap 1: State-of-the-art complexity only applies to **Equality** constraints and **Smooth** Problems

Gap 2: How to efficiently handle **Many** constraints

Turing Inequality into Equality

$$\begin{aligned} \min_{x} \quad & f_0(x) \\ s.t. \quad & g_i(x) \leq 0, i = 1, \ldots, m \end{aligned}$$

$$\begin{aligned} \min_{x,\mathbf{s}} \quad & f_0(x) \\ s.t \quad & g_i(x) + s_i = 0, s_i \geq 0, \forall i \end{aligned}$$

Conversion requires boundness of constraints

Li et al. 2021

$$\begin{aligned} \min_{x,\mathbf{s}} \quad & f_0(x) \\ s.t \quad & g_i(x) + s_i^2 = 0, \forall i \end{aligned}$$

Conversion requires second-order condition of constraints

Ding & J Wright 2023

Our results

- ***Smooth obj & constraints***
  - Single-loop, momentum-style, constraint sampling
  - State-of-the-art

$$O\left(\frac{1}{\epsilon^5}\right)$$

- ***Weakly convex obj & constraints***
  - Double-loop, momentum-style, constraint sampling
  - State-of-the-art

Theory

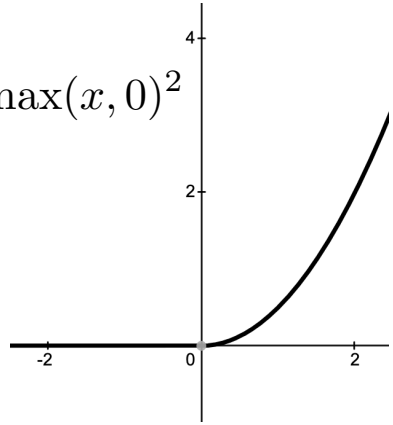
Penalty Methods

penalty func.

$$\min_x f_0(x) + \frac{\rho}{m} \sum_{k=1}^{m} f(g_k(x))$$

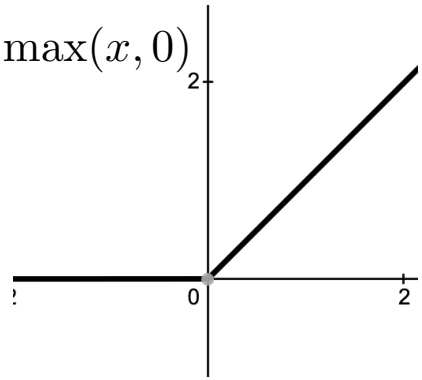
Square hinge penalty

$$f(\cdot) = \frac{1}{2}\max(x, 0)^2$$



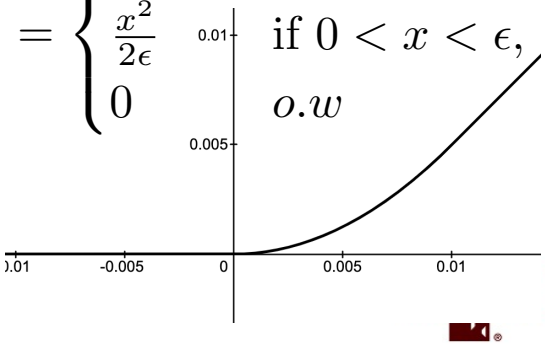
Hinge penalty

$$f(\cdot) = \max(x, 0)$$



Smoothed Hinge penalty

$$f(x) = \begin{cases} x - \frac{\epsilon}{2} & \text{if } x \geq \epsilon, \\ \frac{x^2}{2\epsilon} & \text{if } 0 < x < \epsilon, \\ 0 & o.w \end{cases}$$



Theory

- Squared Hinge Penalty

$$\left\| \nabla f_0(x) + \frac{\rho}{m} \sum_{k=1}^{m} \nabla g_k(x) [g_k(x)]_+ \right\|_2^2 \leq \epsilon^2$$

Regularity condition

$$\sigma_{\min}(\nabla \boldsymbol{g}(x)) \geq \delta, \forall \max_k g_k(x) > 0$$

$$\rho = O\left(\frac{m}{\epsilon\delta^2}\right)$$

epsilon-KKT solution

Theory

- Hinge Penalty

$$\|x - \bar{x}\|_2 \leq O(\epsilon)$$

$$\text{dist}\left(0, \partial f_0(\bar{x}) + \frac{\rho}{m}\sum_{k=1}^{m} \partial f(g_k(\bar{x}))\right) \leq \epsilon$$

Regularity condition

$$\text{dist}\left(0, \frac{1}{m}\sum_{k=1}^{m} \partial [g_k(x)]_+\right) \geq \frac{\delta}{m}, \forall x \in \mathcal{V}$$

$$\rho = O\left(\frac{m}{\delta}\right)$$

Nearly epsilon-KKT solution

Theory

- Smoothed Hinge Penalty

$$\|x - \bar{x}\|_2 \leq O(\epsilon)$$

$$\text{dist}\left(0, \partial f_0(\bar{x}) + \frac{\rho}{m}\sum_{k=1}^{m} \partial f(g_k(\bar{x}))\right) \leq \epsilon$$

Regularity condition

$$\text{dist}\left(0, \partial g(x)^\top \mathbf{v}\right) \geq \delta\|\mathbf{v}\|_2, \forall x \in \mathcal{V}, \forall \mathbf{v} \in \mathbb{R}^m$$

$$\rho = O\left(\frac{m}{\delta}\right)$$

Nearly epsilon-KKT solution

Comparison

	Square hinge penalty	Hinge penalty	Smoothed Hinge penalty
smoothness			
Lipchitz			
Constant penalty			

Algorithms

Challenges

$$\min_x f_0(x) + \frac{\rho}{m} \sum_{k=1}^{m} f(g_k(x))$$

$$\nabla f(g_k(x)) \nabla g_k(x)$$

biased

$$\nabla f(g_k(x; \xi_k')) \nabla g_k(x; \xi_k)$$

Finite-sum Coupled Compositional Optimization (FCCO)

Wang et al. (ICML '22), Jiang et al (NeurIPS '22), Hu et al. (NeurIPS '23), Wang & Yang (ICML'25), Chen et al (NeurIPS '25)

Stochastic Algorithm: smooth case

$$\nabla f(g_k(x))\nabla g_k(x) \qquad k = 1, \ldots, m$$

$$u_{k,t+1} = \begin{cases} (1-\gamma)u_{k,t} + \gamma g_k(x_t;\xi_{k,t}) + \gamma'(g_k(x_t;\xi_{k,t}) - g_k(x_{t-1};\xi_{k,t})) & k = k_t \\ u_{k,t} & \text{o.w} \end{cases}$$

MSVR

$$\gamma' = \frac{n-1}{1-\gamma} + 1 - \gamma$$

Stochastic Algorithm: smooth case

$$\min_x f_0(x) + \frac{\rho}{m}\sum_{k=1}^{m} f(g_k(x))$$

$$z_t = \nabla f_0(x_t;\zeta_t) + \rho \nabla f(u_{k_t,t+1})\nabla g_{k_t}(x_t;\xi'_{k_t})$$

$$m_{t+1} = (1-\beta)m_t + \beta z_t$$

$$O\left(\frac{1}{\epsilon^5}\right)$$

Stochastic Algorithm: non-smooth case

$$\min_x f_0(x) + \frac{\rho}{m}\sum_{k=1}^{m} f(g_k(x))$$

Strongly convex

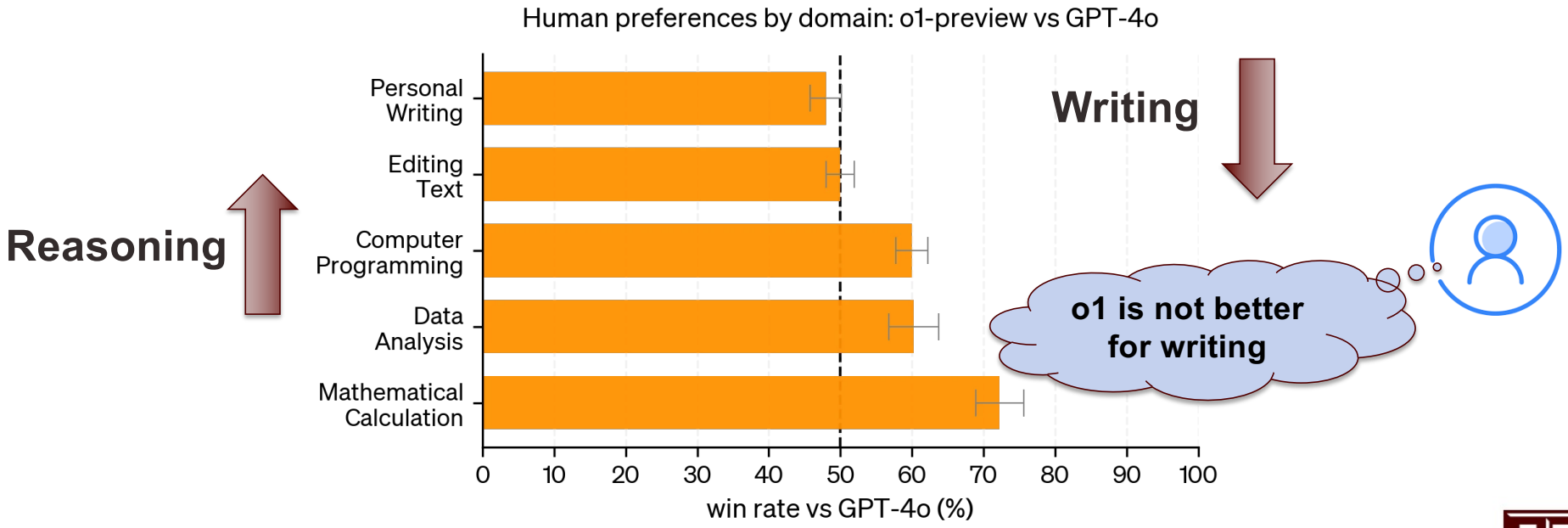
$$\min_x \max_y f_0(x) + \frac{\rho}{m}\sum_{k=1}^{m} y_k g_k(x) - f^*(y_k)$$

Weakly convex → Convex FCCO, Double loop $O\left(\frac{1}{\epsilon^5}\right)$

Experiments

Continual Learning

GPT-3 → GPT-4 → GPT-4o → o1 → o3



Catastrophic Forgetting

Continual learning with Zero-forgetting Constraint

Non-convex Constrained Optimization

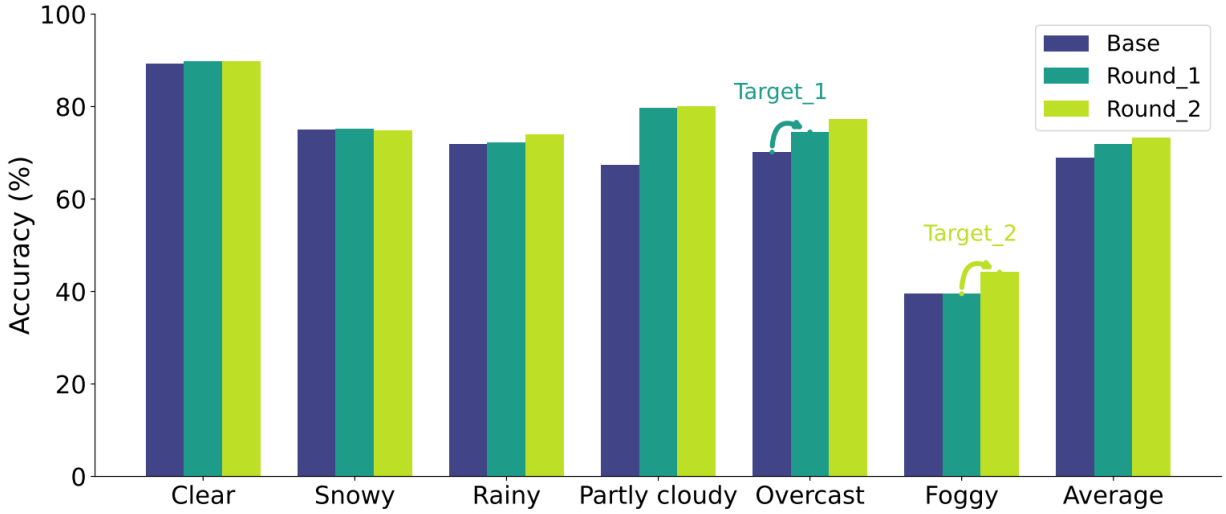
$$\mathbf{w}_{\text{new}} = \arg\min_{\mathbf{w}} F(\mathbf{w})$$

Target Task Objective

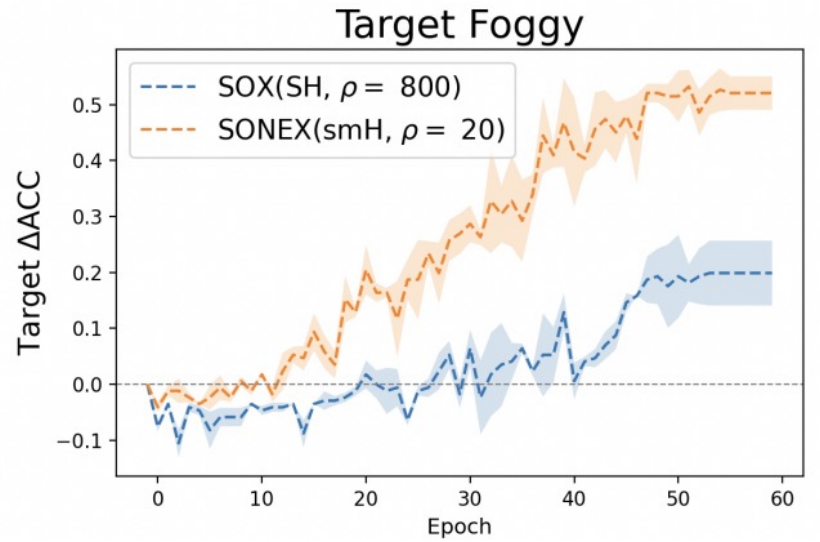
$$s.t. \quad L_k(\mathbf{w}) \leq L_k(\mathbf{w}_{\text{old}}), \quad k = 1, \ldots, m$$

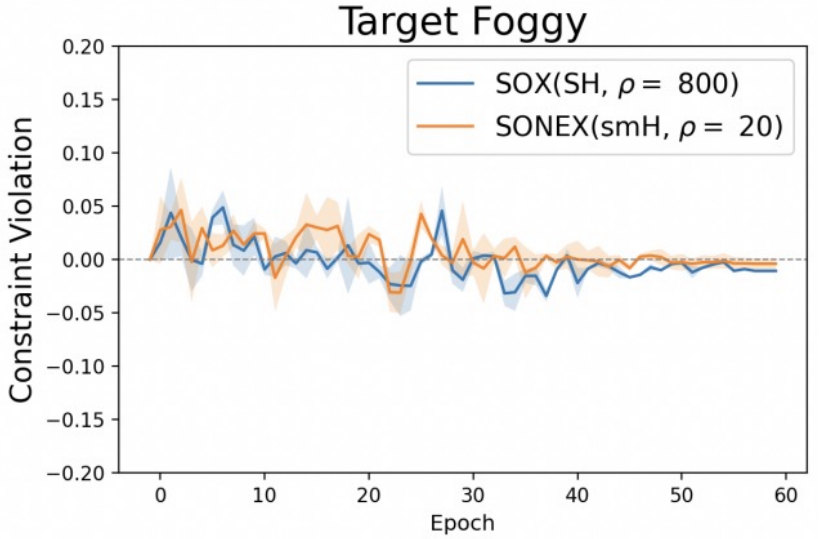
Protected task Loss

Autonomous Driving

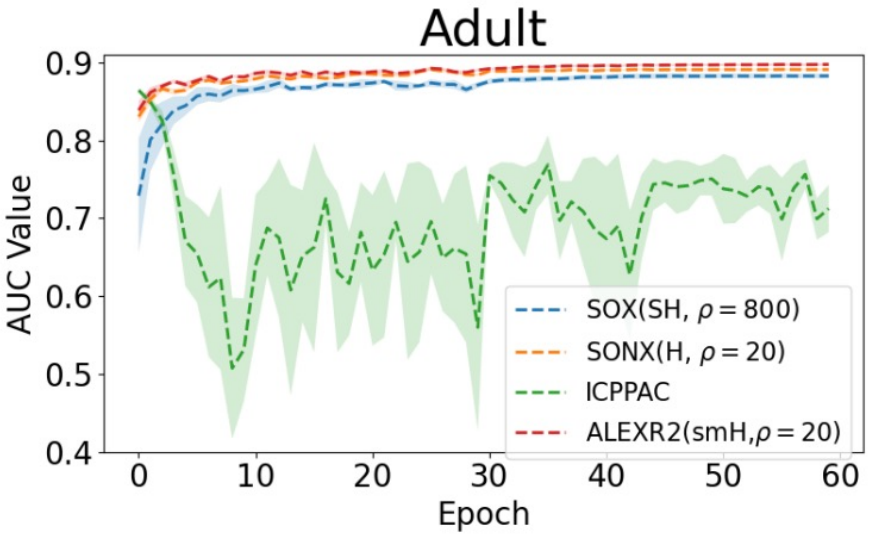


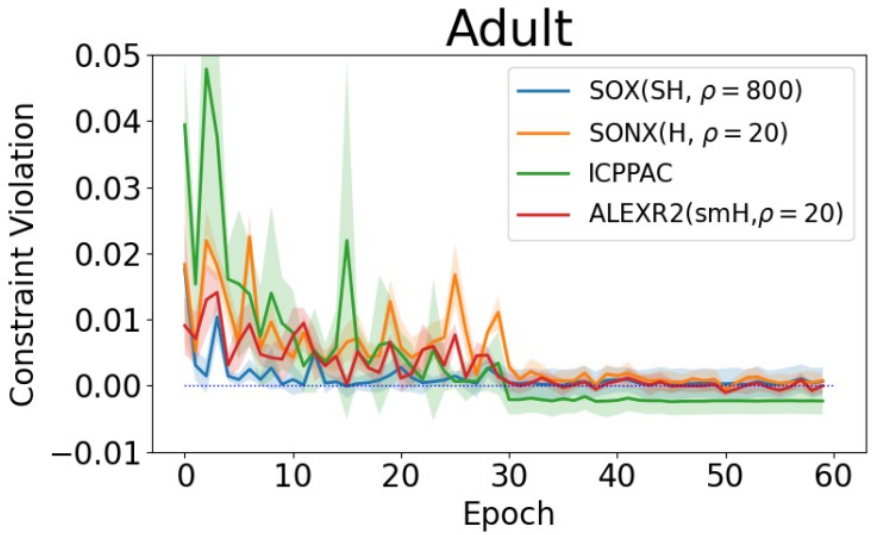
Squared hinge vs Smoothed Hinge





Learning with ROC Fairness





References

Retention-Centric Framework for Continual Learning with Guaranteed Model Developmental Safety. Gang Li et al. arXiv, 2024.

Single-loop Algorithms for Stochastic Non-convex Optimization with Weakly-Convex Constraints. Yang et al. TMLR, 2025.

Stochastic Momentum Methods for Non-smooth Non-Convex Finite-Sum Coupled Compositional Optimization. Chen et al. NeurIPS, 2025.

References

2022
SOX

B. Wang and T. Yang
Finite-Sum Coupled Compositional Stochastic Optimization: Theory and Applications
ICML 2022

2022
MSVR

W. Jiang, G. Li, Y. Wang, L. Zhang, T. Yang
Multi-block-Single-probe Variance Reduced Estimator for Coupled Compositional Optimization
NeurIPS 2022

2023
SONX

Q. Hu, D. Zhu, T. Yang
Non-Smooth Weakly-Convex Finite-sum Coupled Compositional Optimization.
NeurIPS 2023

References

- Ma et al. (’20)
  - *Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints*
- Boob et al. (’23)
  - *Stochastic first-order methods for convex and nonconvex functional constrained optimization*
- Alacaoglu and Wright (’24):
  - *Complexity of single loop algorithms for nonlinear programming with stochastic objective and constraints*
- Li et al. (’24)
  - *Stochastic inexact augmented lagrangian method for nonconvex expectation constrained optimization.*
- Liu & Xu (’25)
  - *A single-loop spider-type stochastic subgradient method for expectation-constrained nonconvex nonsmooth optimization*

Book

Opt4ml.org