Safe Online Convex Optimization with Multi-Point Feedback

Spencer Hutchinson and Mahnoosh Alizadeh

University of California, Santa Barbara

L4DC
2024

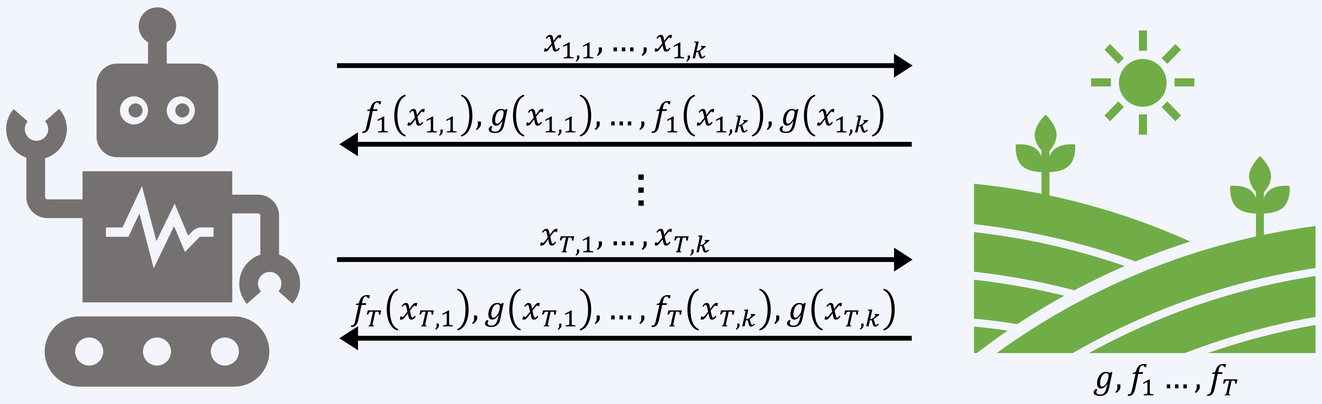
INTRODUCTION

Online convex optimization (OCO) is a powerful theoretical framework that captures a wide variety of sequential decision-making problems. However, most existing OCO approaches *cannot* ensure that unknown round-wise constraints are always satisfied, which is an important requirement for safety-critical applications. To bridge this gap, we study *safe online convex optimization with multi-point feedback*, a version of the OCO problem where the learner must satisfy unknown constraints in every round while only receiving zero-order feedback of costs and constraints. We introduce the algorithm MP-ROGD which enjoys $\mathcal{O}(\sqrt{T})$ regret when the constraint function is smooth and strongly-convex.

Interaction Model:

At each round $t \in [T]$:

1. Player chooses actions $x_{t,1}, \dots, x_{t,k} \in \mathcal{X} \subseteq \mathbb{R}^d$.
2. Adversary chooses cost function f_t.
3. Player incurs cost $\frac{1}{k}\sum_{i=1}^{k} f_t(x_{t,i})$.
4. Player observes multi-point feedback $f_t(x_{t,1}), \dots, f_t(x_{t,k})$, $g(x_{t,1}), \dots, g(x_{t,k})$.



Learning Goals:

- Minimize regret: $R_T = \frac{1}{k}\sum_{t=1}^{T}\sum_{i=1}^{k} f_t(x_{t,i}) - \sum_{t=1}^{T} f_t(x^*)$, $x^* = \operatorname{argmin}_{x\in\mathcal{Y}} \sum_{t=1}^{T} f_t(x)$, $\mathcal{Y} = \{x \in \mathcal{X} : g(x) \le 0\}$
- Satisfy constraint in all rounds: $g(x_t) \le 0 \quad \forall t \in [T]$

Assumptions:

- Action set ($\mathcal{X}$) is convex and has diameter less than D.
- Cost functions (f_t) are convex, L-smooth and gradient norms less than G.
- Constraint function (g) is L-smooth and M-strongly-convex (let $\kappa := \frac{M}{L}$).
- Origin ($\mathbf{0}$) is in the interior of $\mathcal{Y}$, i.e. $r\mathbb{B} \subseteq \mathcal{X}$ and $g(\mathbf{0}) \le -\epsilon$.

ALGORITHM

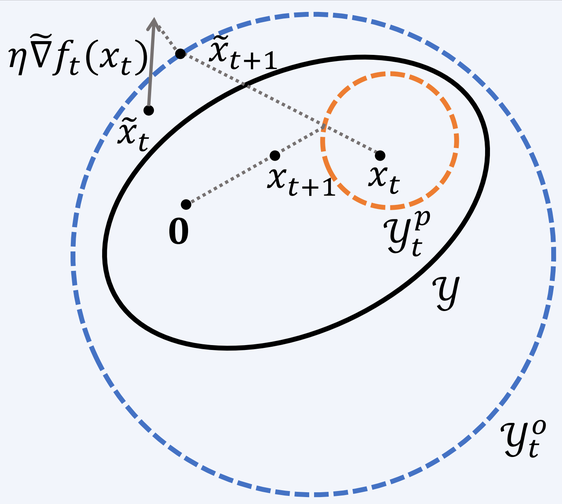
Algorithm 1: Multi-point Restrained Online Gradient Descent (MP-ROGD)

Input: $\mathcal{X}, G, L, M, r, \epsilon, \eta > 0, \delta \in (0,1), \alpha \in (0,1)$.

Set $\tilde{x}_1 = \mathbf{0}$ and $x_1 = \mathbf{0}$.
for $t = 1$ **to** T **do**
  Play $x_t, x_t + \delta e_1, x_t + \delta e_2, \dots, x_t + \delta e_d$.
  Set $\tilde{\nabla} f_t(x_t) = \frac{1}{\delta}\sum_{i=1}^{d}(f_t(x_t + \delta e_i) - f_t(x_t))e_i$ and $\tilde{\nabla} g(x_t) = \frac{1}{\delta}\sum_{i=1}^{d}(g(x_t + \delta e_i) - g(x_t))e_i$.
  Update $\mathcal{Y}_t^o$ and $\mathcal{Y}_t^p$ with (1) and (2).
  $\tilde{x}_{t+1} = \Pi_{\mathcal{Y}_t^o}(\tilde{x}_t - \eta\tilde{\nabla} f_t(x_t))$.
  $\gamma_t = \max\{\mu \in [0,1] : x_t + \mu(\tilde{x}_{t+1} - x_t) \in \mathcal{Y}_t^p\}$.
  $x_{t+1} = (1-\alpha)(x_t + \gamma_t(\tilde{x}_{t+1} - x_t))$.
end

Gradient estimation: The gradients of the cost and constraint functions are estimated with forward-difference. This ensures small gradient estimation error,

$$\left\|\tilde{\nabla} f_t(x_t) - \nabla f_t(x_t)\right\| \le \frac{\sqrt{d}L\delta}{2}.$$



Optimistic and pessimistic sets: The algorithm uses sets that overestimate the feasible set (optimistic set $\mathcal{Y}_t^o$) and underestimate the feasible set (pessimistic set $\mathcal{Y}_t^p$):

$$\mathcal{Y}_t^o = \left\{x \in \mathcal{X} : g(x_t) - \frac{\sqrt{d}L\delta D}{2} + \tilde{\nabla} g(x_t)^\top (x - x_t) + \frac{M}{2}\|x - x_t\|^2 \le 0\right\}$$

$$\mathcal{Y}_t^p = \left\{x \in \mathcal{X} : g(x_t) + \frac{\sqrt{d}L\delta D}{2} + \tilde{\nabla} g(x_t)^\top (x - x_t) + \frac{L}{2}\|x - x_t\|^2 \le 0\right\}$$

From estimation error, smooth and strong-convexity, $\mathcal{Y}_t^p \subseteq \mathcal{Y} \subseteq \mathcal{Y}_t^o$.

Optimistic action: The algorithm updates an optimistic action ($\tilde{x}_t$) using gradient descent over the optimistic set. Since $\mathcal{Y} \subseteq \mathcal{Y}_t^o$, the optimistic action incurs low regret.

Safe step: The algorithm then steps towards the optimistic action while remaining within the pessimistic set. It then steps towards the origin to leave room for the forward-difference gradient estimation.

ANALYSIS

Theorem (Regret Bound)**.** With an appropriate choice of η, δ, α, MP-ROGD ensures that $x_{t,1}, x_{t,2}, \dots, x_{t,k} \in \mathcal{Y}$ for all rounds and that,

$$R_T \le 2DG\sqrt{d\left(\frac{d}{4} + \kappa - 1\right)T} + 1 = \mathcal{O}(d\sqrt{T}).$$

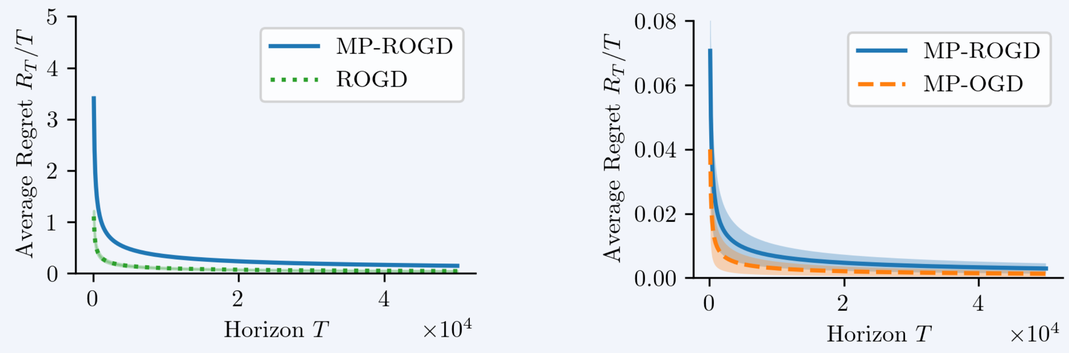
Lemma. With an appropriate choice of α, δ, the played actions stay near to the optimistic actions, i.e.

$$\|x_t - \tilde{x}_t\| \le 2(\kappa - 1)dG\eta = \mathcal{O}\left(\frac{d}{\sqrt{T}}\right).$$

Lemma. With an appropriate choice of α, δ, the optimistic actions incur $\mathcal{O}(d\sqrt{T})$ regret, i.e.

$$\sum_{t=1}^{T} f_t(\tilde{x}_t) - \sum_{t=1}^{T} f_t(x^*) \le \frac{D^2}{2\eta} + \frac{1}{2}d^2G^2\eta T + 1 = \mathcal{O}(d\sqrt{T}).$$

NUMERICAL EXPERIMENTS



FUTURE DIRECTIONS

We will be investigating whether it is possible to get similar guarantees under (1) weaker assumptions on the constraint function, and (2) less constraint feedback. It would also be interesting to see if our algorithmic approach can be applied to related safe learning problems, such as distributed online optimization or online control.

ACKNOWLEDGEMENTS

This work was supported by NSF award #1847096.