Robust Online Convex Optimization in the Presence of Outliers

Tim van Erven

University of Amsterdam

COLT 2021

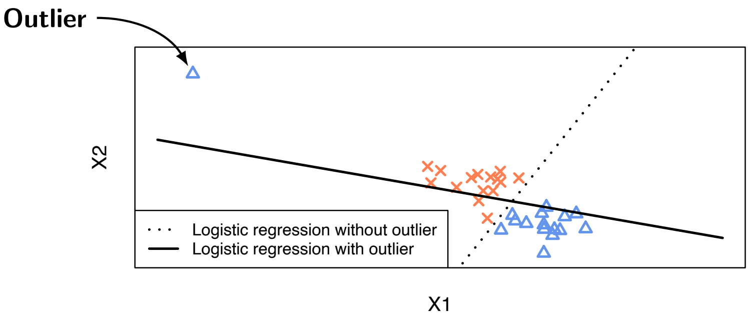
Sarah Sachs

Wouter Koolen

Wojciech Kotłowski

Recruiting: Postdoc position in my group available 2022

Extreme Outliers Can Break Learning



Reasons for outliers:

- Naturally **heavy-tailed data**
- A small subset of **malicious users** trying to corrupt data stream
- Glitches in **cheap sensors**

Heavily studied:

- In statistics [Tukey, 1959, Huber, 1964], stochastic optimization, etc.
- But not yet in Online Convex Optimization

Formalizing Robust OCO

Standard OCO setting:

Given convex domain $\mathcal{W} \subset \mathbb{R}^d$ with diameter$(\mathcal{W}) \leq D$

for $t = 1, 2, \ldots, T$ **do**
    Predict $\boldsymbol{w}_t$ in $\mathcal{W}$
    Observe convex loss function $f_t : \mathcal{W} \to \mathbb{R}$ with gradient $\boldsymbol{g}_t = \nabla f_t(\boldsymbol{w}_t)$
end for

Robust regret: $$R_T(\boldsymbol{u}, \mathcal{S}) = \sum_{t \in \mathcal{S}} \big(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u})\big)$$

Challenges:

- Inliers $\mathcal{S} \subset \{1, \ldots, T\}$ **unknown** (chosen by adversary)
- Bounds **cannot depend on outliers** at all, but must scale with

$$G(\mathcal{S}) = \max_{t \in \mathcal{S}} \|\boldsymbol{g}_t\|.$$

Robustifying Any OCO Algorithm

1. **Any OCO ALG** with regret bound $B_T(G)$ if gradients have length at most G
2. **Top-k Filter**: simple strategy to **filter out large gradients**

Theorem (At most k outliers)

On linear losses, **ALG** + **Top-k Filter** *achieves*

$$R_T(\boldsymbol{u}, \mathcal{S}) \leq \underbrace{B_T\big(2G(\mathcal{S})\big)}_{\textbf{Feed ALG gradients } \leq 2G(\mathcal{S})} + 4DG(\mathcal{S})(k+1) \qquad \textit{for any } \mathcal{S} : T - |\mathcal{S}| \leq k.$$

Robustifying Any OCO Algorithm

1. **Any OCO ALG** with regret bound $B_T(G)$ if gradients have length at most G
2. **Top-k Filter**: simple strategy to **filter out large gradients**

Theorem (At most k outliers)

On linear losses, **ALG** + **Top-k Filter** *achieves*

$$R_T(\boldsymbol{u}, \mathcal{S}) \leq B_T\big(2G(\mathcal{S})\big) + \underbrace{4DG(\mathcal{S})(k+1)}_{\textbf{price of robustness} = O(G(\mathcal{S})k)} \qquad \textit{for any } \mathcal{S} : T - |\mathcal{S}| \leq k.$$

Robustifying Any OCO Algorithm

1. **Any OCO ALG** with regret bound $B_T(G)$ if gradients have length at most G
2. **Top-k Filter**: simple strategy to **filter out large gradients**

Theorem (At most k outliers)

On linear losses, **ALG** + **Top-k Filter** *achieves*

$$R_T(\boldsymbol{u}, \mathcal{S}) \leq B_T\big(2G(\mathcal{S})\big) + 4DG(\mathcal{S})(k+1) \qquad \textit{for any } \mathcal{S} : T - |\mathcal{S}| \leq k.$$

Losses	Minimax Robust Regret
General convex	$O(\sqrt{T} + k)$
General convex + i.i.d.	"
Strongly convex	$O(\ln(T) + k)$

Efficient Filtering Approach

Top-k Filter:

- Maintain list $\mathcal{L}_t$ of $k+1$ largest gradient lengths seen so far
- Filter round if $\|\boldsymbol{g}_t\| > 2 \min \mathcal{L}_t$; otherwise pass to ALG

Main Ideas:

1. Never pass ALG gradients $> 2G(\mathcal{S})$:
   - $\mathcal{L}_t$ contains at least 1 inlier, because at most k outliers
   - Hence $\min \mathcal{L}_t \leq G(\mathcal{S})$
2. Overhead for filtering is $O(k)$
   - Every filtered round is also added to $\mathcal{L}_t$
   - Therefore $\min \mathcal{L}_t$ (at least) doubles every $k+1$ filtered rounds
   - Hence last $k+1$ filtered rounds dominate

Application: Robustified Online-to-Batch

Outlier distribution

Huber ϵ-contamination model: $P_\epsilon = (1-\epsilon)P + \epsilon Q$

Distribution of interest

- $f_t(\boldsymbol{w}) = f(\boldsymbol{w}, \xi)$ where $\xi \sim P_\epsilon$
- Inlier risk: $\mathrm{Risk}_P(\boldsymbol{w}) = \mathbb{E}_{\xi \sim P}[f(\boldsymbol{w}, \xi)]$

Application: Robustified Online-to-Batch

Outlier distribution

Huber ϵ-contamination model: $P_\epsilon = (1-\epsilon)P + \epsilon Q$

Distribution of interest

- $f_t(\boldsymbol{w}) = f(\boldsymbol{w}, \xi)$ where $\xi \sim P_\epsilon$
- Inlier risk: $\text{Risk}_P(\boldsymbol{w}) = \mathbb{E}_{\xi \sim P}[f(\boldsymbol{w}, \xi)]$

Corollary (Optimal Rate via Robust Online-to-Batch)

Suppose $\|\nabla f(\boldsymbol{w}, \xi)\| \le G$ a.s. when $\xi \sim P$ is an inlier.
Then iterate average $\bar{\boldsymbol{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t$ of **OGD** + **Top-k Filter** *achieves*

$$\text{Risk}_P(\bar{\boldsymbol{w}}_T) - \min_{\boldsymbol{u} \in \mathcal{W}} \text{Risk}_P(\boldsymbol{u}) = O\left(DG\epsilon + DG\sqrt{\frac{\ln(1/\delta)}{T}}\right)$$

with **P_ϵ-probability** *at least $1-\delta$, for some k tuned for ϵ, δ, T.*

Quantile Outliers

Which **extra assumptions** allow
sublinear dependence on number of outliers k?

- $\|\boldsymbol{g}_t\| \leq L\|\boldsymbol{X}_t\|$ for i.i.d. $\boldsymbol{X}_t$ (e.g. hinge loss, logistic loss)
- Inliers $\mathcal{S}_p$ are rounds s.t. $\|\boldsymbol{X}_t\|$ less than p-quantile X_p

Quantile Outliers

Which **extra assumptions** allow
sublinear dependence on number of outliers k?

- $\|\boldsymbol{g}_t\| \le L\|\boldsymbol{X}_t\|$ for i.i.d. $\boldsymbol{X}_t$ (e.g. hinge loss, logistic loss)
- Inliers $\mathcal{S}_p$ are rounds s.t. $\|\boldsymbol{X}_t\|$ less than p-quantile X_p

Theorem (Sublinear Outlier Overhead)

Suppose ALG has regret bound $B_T(X)$, concave in T, if non-filtered $\boldsymbol{X}_t$ have length at most X. Then **ALG** + ***p*-Quantile Filter** *achieves*

$$\mathbb{E}\left[\max_{\boldsymbol{u}\in\mathcal{W}} R_T(\boldsymbol{u}, \mathcal{S}_p)\right] \le B_{pT}(X_p) + O\left(LDX_p\sqrt{p(1-p)T\ln T} + \ln(T)^2\right).$$

***p*-Quantile Filter:**

- Filter when $\|\boldsymbol{X}_t\| \ge$ lower-confidence bound on X_p

Summary

Robust regret: measure regret only on (unknown) inlier rounds

Price of Robustness = Overhead over usual regret rate:

- At most k adversarial outliers: $O(k)$
- p-Quantile outliers: $O(\sqrt{p(1-p)T\ln(T)} + \ln(T)^2)$

**PS. I am looking for a postdoc, starting anytime in 2022.
Please get in touch if you want to come to Amsterdam!**