UvA-DARE (Digital Academic Repository)

Robust Online Convex Optimization in the Presence of Outliers

van Erven, T.; Sachs, S.; Koolen, W.M.; Kotłowski, W.

Robust Online Convex Optimization in the Presence of Outliers

Tim van Erven   TIM@TIMVANERVEN.NL
Sarah Sachs   S.C.SACHS@UVA.NL
Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands

Wouter M. Koolen   WMKOOLEN@CWI.NL
Centrum Wiskunde & Informatica, The Netherlands

Wojciech Kotłowski   WKOTLOWSKI@CS.PUT.POZNAN.PL
Poznań University of Technology, Poland

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We consider online convex optimization when a number k of data points are outliers that may be corrupted. We model this by introducing the notion of robust regret, which measures the regret only on rounds that are not outliers. The aim for the learner is to achieve small robust regret, without knowing where the outliers are. If the outliers are chosen adversarially, we show that a simple filtering strategy on extreme gradients incurs $O(k)$ additive overhead compared to the usual regret bounds, and that this is unimprovable, which means that k needs to be sublinear in the number of rounds. We further ask which additional assumptions would allow for a linear number of outliers. It turns out that the usual benign cases of independently, identically distributed (i.i.d.) observations or strongly convex losses are not sufficient. However, combining i.i.d. observations with the assumption that outliers are those observations that are in an extreme quantile of the distribution, does lead to sublinear robust regret, even though the expected number of outliers is linear.

Keywords: Online convex optimization, robustness, outliers

1. Introduction

Methods for online convex optimization (OCO) are designed to work even in the presence of adversarially generated data (Hazan, 2015; Shalev-Shwartz, 2011; Cesa-Bianchi and Lugosi, 2006), but this is only possible because strong boundedness assumptions are imposed on the losses that limit the influence of individual data points. On the other hand, the most practically successful methods do not enforce an a priori specified bound on the losses, but instead adapt to the norms of the observed gradients or to the observed loss range. For example, the regret bound for AdaGrad adapts to the ranges of the gradient components per dimension (Duchi et al., 2011), the regret bound for online ridge regression scales with the largest observed loss (Vovk, 2001), the regret bound for AdaHedge in the prediction with experts setting scales with the observed loss range of the experts (De Rooij et al., 2014), the regret bound for online gradient descent on strongly convex losses scales with the maximum gradient norm squared (Hazan et al., 2007), etc. In all such cases a small number of outliers with large gradients among an otherwise benign dataset can significantly worsen performance. This is also clear directly from the algorithms themselves, where we see that large gradients have the effect of significantly decreasing the effective step size for all subsequent data points, leading to slower learning. Extreme outliers may occur naturally, for instance because of heavy-tailed distributions or sensor glitches, but if each loss is based on the input of a user, then we may also

be concerned that a small number of adversarial users may try to poison the data stream (Kurakin et al., 2017).

We formally capture the robustness of OCO methods by modifying the standard setting to measure performance only on the rounds that are not outliers. The goal of the learner is to perform as well as if the outliers were not present, up to some overhead that is incurred for filtering out the outliers. As in standard OCO, learning proceeds in T rounds, and at the start of each round t the learner needs to issue a prediction $\boldsymbol{w}_t$ from a bounded convex domain. The environment then reveals a convex loss function f_t with (sub)gradient $\boldsymbol{g}_t := \nabla f_t(\boldsymbol{w}_t)$ at $\boldsymbol{w}_t$, and performance is measured by the cumulative difference between the learner's losses and the losses of the best fixed parameters $\boldsymbol{u}$. Unlike in the standard OCO setting, however, we only sum up losses over the subset of inlier rounds $\mathcal{S} \subseteq \{1, 2, \ldots, T\}$ that are not outliers, leading to the following notion of *robust regret*:

$$R_T(\boldsymbol{u}, \mathcal{S}) := \sum_{t \in \mathcal{S}} \big(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u})\big). \tag{1}$$

The challenge for the learner is to guarantee small robust regret without knowing $\mathcal{S}$. Importantly, we aim for robust regret bounds that scale with the loss range or gradient norms of the rounds in $\mathcal{S}$, but not with the size of the outliers, so even extreme outliers should not be able to confuse the learner.

In Section 3, we first consider the fully adversarial case where the only thing the learner knows is that there are at most k outliers, so $T - |\mathcal{S}| \leq k$, and both the inliers and the outliers are generated adversarially, without any bound on the range of the outliers, and with the range of the inliers also unknown a priori. We introduce a simple filtering approach that filters out some of the largest gradients, and passes on the remaining rounds to a standard online learning algorithm ALG. When the losses are linear, this approach is able to guarantee that

$$R_T(\boldsymbol{u}, \mathcal{S}) = R_T^{\text{ALG}}(\boldsymbol{u}) + O\big(G(\mathcal{S})k\big) \qquad \text{for all } \mathcal{S} \text{ such that } T - |\mathcal{S}| \leq k \text{ simultaneously,} \tag{2}$$

where $G(\mathcal{S})$ is the norm of the largest gradient among the rounds in $\mathcal{S}$ and $R_T^{\text{ALG}}(\boldsymbol{u})$ is the regret of ALG on a subset of rounds under the guarantee that their gradient norms are at most $2G(\mathcal{S})$. The extension to general convex losses then follows from a standard reduction to the linear loss case. We follow up by showing that (2) is unimprovable, not just for adversarial losses, but even if the losses are independent and identically distributed (i.i.d.) according to a fixed probability distribution or if the losses are strongly convex. This fixes the dependence on the number of outliers k to be linear in k in quite some generality. Nevertheless, in Section 4 we identify sufficient conditions to get around the linear dependence: if the gradients are i.i.d., and we take $\mathcal{S} = \mathcal{S}_p$ to be the rounds in which $\|\boldsymbol{g}_t\|_*$ is at most the p-quantile G_p of the common distribution of their norms, then there exists a method based on approximating G_p by its empirical counterpart on the available data that guarantees that the expected robust regret is at most

$$\mathbb{E}\left[R_T(\boldsymbol{u}, \mathcal{S}_p)\right] = O\left(G_p\left(\sqrt{pT} + \sqrt{p(1-p)T \ln T} + \ln^2 T\right)\right). \tag{3}$$

Since $O\big(G_p \sqrt{pT}\big)$ would be expected if $\mathcal{S}_p$ were known in advance, we see that the overhead grows sublinearly in T and is even asymptotically negligible for outlier proportion $1 - p = o(1/\ln(T))$. More generally, we extend this result such that the gradients do not need to be i.i.d. themselves, but it is sufficient if there exist i.i.d. random variables $\boldsymbol{X}_t$ and a constant L such that $\|\boldsymbol{g}_t\|_* \leq L\|\boldsymbol{X}_t\|_*$. We then define the quantile with respect to the distribution of the $\boldsymbol{X}_t$. This covers nonlinear losses

of the form $f_t(\boldsymbol{w}) = h_t(\boldsymbol{w}^\intercal \boldsymbol{X}_t)$ for convex functions h_t that are L-Lipschitz, like the logistic loss $f_t(\boldsymbol{w}) = \ln(1 + \exp(-Y_t \boldsymbol{w}^\intercal \boldsymbol{X}_t))$ and the hinge loss $f_t(\boldsymbol{w}) = \max\{1 - Y_t \boldsymbol{w}^\intercal \boldsymbol{X}_t\}$ for $Y_t \in \{-1, +1\}$, both with $L = 1$.

Related Work   The definition of robust regret may remind the reader of the adaptive regret (Hazan and Seshadhri, 2009; Daniely et al., 2015), which measures regret on a contiguous interval of rounds $\mathcal{I}$ that is unknown to the learner. Since adaptive regret can be controlled by casting it into the framework of specialist (sleeping) experts (Freund et al., 1997; Chernov and Vovk, 2009), it is natural to ask whether the same is possible for the robust regret. To apply the specialist experts framework, we would assign a separate learner (specialist) to each possible subset of rounds $\mathcal{S}$ that would then be active only on $\mathcal{S}$, and such a pool of m learners would be aggregated using a meta-algorithm. Computational issues aside, this approach runs into two problems: the first is that all existing meta-algorithms assume the losses to be bounded within a known range, and therefore cannot be applied since we do not assume that even the range of the inliers is known. Second, even if the range issue could be resolved, the specialist regret would incur a $\Omega(\sqrt{T \log m}) = \Omega(\sqrt{kT \log(T/k)})$ overhead, already if we only consider all $m = \binom{T}{T-k} \geq (T/k)^k$ possible subsets with exactly k outliers. We see that k now multiplies T, which is much worse than the optimal additive dependence on k in (2).

Reducing the dependence on the largest gradient norm has previously been considered in the context of adaptive online and stochastic convex optimization (Duchi et al., 2011; Ward et al., 2019). However, these methods still depend on the average of all (squared) gradient norms, and therefore require these norms to be finite. In contrast to these adaptive methods, our method can handle a small number of adversarial samples, with large or even infinite norm, while our robust regret analysis still guarantees a sub-linear bound.

In the context of stochastic optimization, Nazin et al. (2019) propose a robust version of mirror descent based on truncating the gradients returned by a stochastic oracle. Their main goal is to establish a sub-Gaussian confidence bound on the optimization error under weak assumptions about the tails of the noise distribution. Contrary to our setup, they control the smoothness of the objective and the variance of the noise, so that already a vanilla (non-robust) version of SGD would achieve a vanishing optimization error in expectation (but not with a sub-Gaussian confidence). Diakonikolas et al. (2019a) propose a robust meta-algorithm for stochastic optimization that repeatedly trains a standard algorithm as a base learner and filters out the outliers. This approach is conceptually similar to our filtering method, but it is designed to work in a batch setting, with the data (sample functions) given in advance. Prasad et al. (2018) provide a robust batch algorithm for stochastic optimization by applying the ideas from robust mean estimation to robustify stochastic gradient estimates in a (batch) gradient descent algorithm.

In the online learning and bandit literature, interesting results were obtained for dealing with adversarial corruptions of data that are otherwise generated i.i.d., to still benefit from the stochastic setting (Lykouris et al., 2018; Gupta et al., 2019; Amir et al., 2020). Wang and Chaudhuri (2018) and Zhang et al. (2020) further consider data poisoning attacks on an online learner, but the focus is on the optimization of the adversary, while the learner remains fixed. In all these works, contrary to ours, the corrupted data is still assumed to lie in the same range as the non-corrupted data. A notable exception is the very recent work of Chen et al. (2020), which proposes online algorithms for contextual bandits and linear regression in a framework in which the linear model is realizable (well-specified) up to small noise, and a fixed, randomly selected, fraction of examples is arbitrarily corrupted (as in the Huber ϵ-contamination model (Huber, 1964)), but still remains bounded. In

contrast, we avoid strong distributional assumptions such as model realizability, and do not make any probabilistic assumptions about the corruption mechanism or impose any constraints on the magnitude of the outliers.

Starting with pioneering works of Tukey and Huber (Tukey, 1959; Huber, 1964) there has been a tremendous amount of past work in the area of robust statistics, which concerns the basic tasks of classical statistics in the presence of outliers and heavy-tailed distributions (Huber, 1981). A more recent line of research building on the work of Catoni (2012); Minsker (2015); Lugosi and Mendelson (2019b,a) concerns estimation with sub-Gaussian-style confidence for heavy-tailed distributions. Finally, our setup is different from, but conceptually related to, a line of research on machine learning and statistical problems in the presence of adversarial data corruptions (Charikar et al., 2017). This has been studied, for instance, in the context of parameter estimation (Lai et al., 2016; Cheng et al., 2019; Diakonikolas et al., 2018; Lugosi and Mendelson, 2019c; Prasad et al., 2020), robust PCA (Candès et al., 2011), regression (Klivans et al., 2018; Diakonikolas et al., 2019b; Liu et al., 2020), classification (Klivans et al., 2009; Liu and Tao, 2016; Awasthi et al., 2017) and many other cases. See the in-depth survey by Diakonikolas and Kane (2019) for an overview of recent advances in this direction.

Outline We start by summarizing our setting and notation in the next section. Then, in Section 3, we prove the upper bound (2) for adversarial losses, and show matching lower bounds both for i.i.d. losses and for strongly convex losses. As a further example, we show how robust regret can be used to bound the excess risk in the Huber ϵ-contaminated setting via online-to-batch-conversion. In Section 4 we turn to the quantile case and establish (3). Finally, Section 5 concludes with a discussion of possible directions for future work. Some proofs are deferred to the appendix.

2. Setting and Notation

Formally, we consider the following online learning protocol. In each round $t = 1, 2, \ldots$ the learner first predicts $\boldsymbol{w}_t \in \mathcal{W}$, where the domain $\mathcal{W}$ is a non-empty, compact and convex subset of $\mathbb{R}^d$. The adversary then reveals a convex loss function $f_t : \mathcal{W} \to \mathbb{R}$, and the learner suffers loss $f_t(\boldsymbol{w}_t)$. We assume throughout that there always exists a gradient or, more generally, a subgradient $\boldsymbol{g}_t := \nabla f_t(\boldsymbol{w}_t)$ at the learner's prediction, which is implied by convexity of f_t whenever $\boldsymbol{w}_t$ lies in the interior of $\mathcal{W}$ and also on the boundary if there exists a finite convex extension of all f_t to a larger domain that contains $\mathcal{W}$ in its interior. The performance of the learner with respect to any fixed parameters $\boldsymbol{u} \in \mathcal{W}$ is measured by the *robust regret* $R_T(\boldsymbol{u}, \mathcal{S})$ over the rounds $\mathcal{S} \subseteq [T] := \{1, 2, \ldots, T\}$ that are not outliers, as defined in (1). The definition of subgradients implies that $f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u}) \le (\boldsymbol{w}_t - \boldsymbol{u})^\mathsf{T} \boldsymbol{g}_t$, which implies that $R_T(\boldsymbol{u}, \mathcal{S})$ is bounded from above by the *linearized robust regret*

$$\widetilde{R}_T(\boldsymbol{u}, \mathcal{S}) := \sum_{t \in \mathcal{S}} (\boldsymbol{w}_t - \boldsymbol{u})^\mathsf{T} \boldsymbol{g}_t.$$

We will state our main results for an arbitrary norm $\|\cdot\|$ on $\mathcal{W}$ and measure gradient lengths in terms of the dual norm $\|\boldsymbol{g}_t\|_* = \sup_{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \le 1} \boldsymbol{w}^\mathsf{T} \boldsymbol{g}_t$. Let $D = \max_{\boldsymbol{u}, \boldsymbol{w} \in \mathcal{W}} \|\boldsymbol{w} - \boldsymbol{u}\|$ denote the diameter of the domain. For the analysis of the robust regret, we need a Lipschitz bound for the gradients that are in the set $\mathcal{S}$, which we denote by

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

3. Robustness to Adversarial Outliers

In this section we derive matching upper and lower bounds of the form in (2).

3.1. Upper Bounds

Let ALG be any Lipschitz-adaptive algorithm, which we will use as our base online learning algorithm. Our general approach is to add a filtering meta-algorithm FILTER that examines (the norm of) incoming gradients and decides whether to filter them or pass them on to ALG for learning. If $\mathcal{S}$ were known in advance, then FILTER could filter out all outliers and pass on only the rounds in $\mathcal{S}$, but since $\mathcal{S}$ is not known, FILTER needs to learn which rounds to pass on. Although most online learning algorithms base their updates only on gradients, we note that we do allow ALG to use the full loss function f_t to update its state when FILTER passes on round t to ALG. When a round t is filtered, we assume that ALG behaves as if that round had not happened, so we will have $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$. Our FILTER for this section is displayed in Algorithm 1.

Algorithm 1: Top-k Filter: Filtering for Adversarial Setting

Input: Maximum number of outliers k
Initialize: Let $\mathcal{L}_0 = \{0, 0, \ldots, 0\}$ be an ordered list of length $k + 1$.
for $t = 1, 2, \ldots$ **do**

    Maintain invariant that $\mathcal{L}_t$ contains $k+1$ largest gradients;
    if $\|\boldsymbol{g}_t\|_* > \min \mathcal{L}_{t-1}$ **then**
        Obtain $\mathcal{L}_t$ from $\mathcal{L}_{t-1}$ by removing the smallest item in $\mathcal{L}_{t-1}$ and inserting $\|\boldsymbol{g}_t\|_*$;
    else
        Set $\mathcal{L}_t$ equal to $\mathcal{L}_{t-1}$;
    end
    Filter with factor 2 slack;
    if $\|\boldsymbol{g}_t\|_* > 2 \min \mathcal{L}_t$ **then**
        Filter round t;
    else
        Pass round t on to ALG;
    end
end

Theorem 1 *Suppose ALG is any Lipschitz-adaptive algorithm that guarantees linearized regret bounded by $B_T(G)$ on the rounds that it is passed by FILTER, if the gradients in those rounds have length at most G, and let FILTER be Algorithm 1 with parameter k. Then the linearized robust regret of ALG+FILTER is bounded by*

$$\widetilde{R}_T(\boldsymbol{u}, \mathcal{S}) \leq B_T\big(2G(\mathcal{S})\big) + 4D(\boldsymbol{u}, \mathcal{S})G(\mathcal{S})(k+1) \qquad \text{for any } \mathcal{S} : T - |\mathcal{S}| \leq k \text{ and } \boldsymbol{u} \in \mathcal{W}, \quad (4)$$

where $D(\boldsymbol{u}, \mathcal{S}) = \max_{t:\|\boldsymbol{g}_t\|_* \leq 2G(\mathcal{S})} \|\boldsymbol{w}_t - \boldsymbol{u}\|$.

There are two main ideas to the proof. First, since the list $\mathcal{L}_t$ in Algorithm 1 contains $k+1$ elements and there are at most k outliers, at least one of the elements of $\mathcal{L}_t$ must be one of the inliers from $\mathcal{S}$.

It follows that the smallest element of $\mathcal{L}_t$ is a lower bound on $G(\mathcal{S})$. The second idea is that, instead of filtering on this lower bound directly, we filter with factor 2 slack. Since every filtered gradient is also added to $\mathcal{L}_t$, this factor 2 ensures that the minimum of $\mathcal{L}_t$ must at least double for every $k + 1$ rounds that are filtered. The resulting exponential growth of the filtered rounds means that the contribution to the robust regret of all filtered rounds is dominated by the last $k + 1$ rounds, and therefore does not grow with T.

Proof Let $\mathcal{F} \subset [T]$ denote the rounds filtered out by Algorithm 1, and let $\mathcal{P} = [T] \setminus \mathcal{F}$ denote the rounds that are passed on to ALG. Then the linearized robust regret splits as follows:

$$\widetilde{R}_T(\boldsymbol{u}, \mathcal{S}) = \sum_{t \in \mathcal{S} \cap \mathcal{P}} (\boldsymbol{w}_t - \boldsymbol{u})^\mathsf{T} \boldsymbol{g}_t + \sum_{t \in \mathcal{S} \cap \mathcal{F}} (\boldsymbol{w}_t - \boldsymbol{u})^\mathsf{T} \boldsymbol{g}_t.$$

We will show that Algorithm 1 guarantees that the gradients on the passed rounds are bounded as follows:

$$\|\boldsymbol{g}_t\|_* \leq 2G(\mathcal{S}) \qquad \text{for all } t \in \mathcal{P}, \tag{5}$$

which implies that

$$\sum_{t \in \mathcal{S} \cap \mathcal{P}} (\boldsymbol{w}_t - \boldsymbol{u})^\mathsf{T} \boldsymbol{g}_t = \sum_{t \in \mathcal{P}} (\boldsymbol{w}_t - \boldsymbol{u})^\mathsf{T} \boldsymbol{g}_t - \sum_{t \in \mathcal{P} \setminus \mathcal{S}} (\boldsymbol{w}_t - \boldsymbol{u})^\mathsf{T} \boldsymbol{g}_t$$

$$\leq B_T(2G(\mathcal{S})) + 2D(\boldsymbol{u}, \mathcal{S})G(\mathcal{S})|\mathcal{P} \setminus \mathcal{S}| \leq B_T(2G(\mathcal{S})) + 2D(\boldsymbol{u}, \mathcal{S})G(\mathcal{S})k,$$

where the first inequality uses the assumption on ALG and Hölder's inequality, and the second inequality uses that $|\mathcal{P} \setminus \mathcal{S}| \leq |[T] \setminus \mathcal{S}| \leq k$.

We proceed to prove (5). During the first k rounds, $\min \mathcal{L}_t = 0$, so (5) is trivially satisfied. In all later rounds, $\mathcal{L}_t \subseteq \{\|\boldsymbol{g}_s\|_* : s \leq t\} \subseteq \{\|\boldsymbol{g}_s\|_* : s \leq T\}$. Consequently, $\mathcal{L}_t$ must contain at least one element $\|\boldsymbol{g}_t\|_*$ with $t \in \mathcal{S}$, because $T - |\mathcal{S}| \leq k$ and $|\mathcal{L}_t| = k + 1 > k$. It follows that $\min \mathcal{L}_t \leq G(\mathcal{S})$, so all passed gradients satisfy (5).

Let $G_{\min} = \min\{\|\boldsymbol{g}_t\|_* \mid t \in \mathcal{F}\} > 0$ be the length of the shortest filtered gradient. To complete the proof, we will show that

$$\sum_{t \in \mathcal{S} \cap \mathcal{F}} (\boldsymbol{w}_t - \boldsymbol{u})^\mathsf{T} \boldsymbol{g}_t \leq D(\boldsymbol{u}, \mathcal{S}) \sum_{t \in \mathcal{S} \cap \mathcal{F}} \|\boldsymbol{g}_t\|_* \leq D(\boldsymbol{u}, \mathcal{S}) \sum_{\substack{t \in \mathcal{F} \\ G_{\min} \leq \|\boldsymbol{g}_t\|_* \leq G(\mathcal{S})}} \|\boldsymbol{g}_t\|_* \leq 2D(\boldsymbol{u}, \mathcal{S})G(\mathcal{S})(k + 1).$$

The first of these inequalities follows from Hölder's inequality, and the second from the definition of $G(\mathcal{S})$. To establish the last inequality, we proceed by induction: since $\mathcal{L}_t$ contains the $k + 1$ largest observed gradient norms, we observe that there can be at most $k + 1$ filtered rounds in which $G(\mathcal{S})/2^{i+1} < \|\boldsymbol{g}_t\|_* \leq G(\mathcal{S})/2^i$, because after $k + 1$ such rounds we will have $\min \mathcal{L}_t > G(\mathcal{S})/2^{i+1}$ forever. It follows that we have the following induction step:

$$\sum_{\substack{t \in \mathcal{F} \\ G_{\min} \leq \|\boldsymbol{g}_t\|_* \leq G(\mathcal{S})/2^i}} \|\boldsymbol{g}_t\|_* \leq (k + 1)G(\mathcal{S})/2^i \quad + \sum_{\substack{t \in \mathcal{F} \\ G_{\min} \leq \|\boldsymbol{g}_t\|_* \leq G(\mathcal{S})/2^{i+1}}} \|\boldsymbol{g}_t\|_* \qquad \text{for } i = 0, 1, 2, \ldots$$

Unrolling the induction, we therefore obtain

$$\sum_{\substack{t \in \mathcal{F} \\ G_{\min} \leq \|\boldsymbol{g}_t\|_* \leq G(\mathcal{S})}} \|\boldsymbol{g}_t\|_* \leq (k + 1)G(\mathcal{S}) \sum_{i=0}^{\lceil \log_2 \frac{G(\mathcal{S})}{G_{\min}} \rceil} 2^{-i} \leq (k + 1)G(\mathcal{S}) \sum_{i=0}^{\infty} 2^{-i} = (k + 1)G(\mathcal{S})2,$$

which is what remained to be shown. ■

As for the run-time, one may maintain the k largest gradient norms encountered in a priority queue. The time used by Algorithm 1 on top of ALG is $O(\ln k) \leq O(\ln T)$ per round. This may be pessimistic in practise, as FILTER only performs work if the current gradient is among the $k + 1$ largest seen so far.

3.1.1. EXAMPLES

To make the result from Theorem 1 more concrete, let us instantiate ALG as online gradient descent (OGD), which starts from any $\boldsymbol{w}_1 \in \mathcal{W}$ and updates according to

$$\boldsymbol{w}_{t+1} = \Pi_{\mathcal{W}}(\boldsymbol{w}_t - \eta_t \boldsymbol{g}_t),$$

where $\Pi_{\mathcal{W}}(\boldsymbol{w})$ denotes Euclidean projection of $\boldsymbol{w}$ onto $\mathcal{W}$, and $\eta_t > 0$ is a hyperparameter called the step size. Tuning the step size for general convex losses, we find that we can tolerate at most $k = O(\sqrt{T})$ outliers without suffering in the rate:

Corollary 2 (General Convex Losses) *Let $\|\cdot\|$ be the ℓ_2-norm, let ALG be OGD with step size $\eta_t = D/\sqrt{2\sum_{s=1}^{t}\|\boldsymbol{g}_s\|_2^2}$ and let FILTER be Algorithm 1 with parameter k. Then the robust regret is bounded by*

$$
\begin{aligned}
R_T(\boldsymbol{u}, \mathcal{S}) &\leq 2D\sqrt{\sum_{t\in\mathcal{S}}\|\boldsymbol{g}_t\|_2^2} + 2DG(\mathcal{S})\Big(2k + \sqrt{k} + 2\Big) \\
&\leq 2DG(\mathcal{S})\Big(\sqrt{T} + 2k + \sqrt{k} + 2\Big) \quad \textit{for any } \mathcal{S}: T - |\mathcal{S}| \leq k \textit{ and } \boldsymbol{u} \in \mathcal{W}.
\end{aligned}
\tag{6}
$$

The proof of the corollary is in Appendix A.

The step size of OGD may also be tuned for *σ-strongly convex losses*, which are guaranteed to be curved in all directions, and satisfy the requirement that

$$f_t(\boldsymbol{u}) \geq f_t(\boldsymbol{w}) + (\boldsymbol{u} - \boldsymbol{w})^{\mathsf{T}}\nabla f_t(\boldsymbol{w}) + \frac{\sigma}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2 \qquad \text{for all } \boldsymbol{u}, \boldsymbol{w} \in \mathcal{W}.$$

In this case, we obtain the following guarantee on the robust regret, which is proved in Appendix A:

Corollary 3 (Strongly Convex Losses) *Suppose the loss functions f_t are σ-strongly convex. Let $\|\cdot\|$ be the ℓ_2-norm, let ALG be OGD with step size $\eta_t = \frac{1}{\sigma t}$ and let FILTER be Algorithm 1 with parameter k. Then the robust regret is bounded by*

$$R_T(\boldsymbol{u}, \mathcal{S}) \leq \frac{2G(\mathcal{S})^2}{\sigma}\big(\ln T + 1\big) + \frac{5\widetilde{G}(\boldsymbol{u}, \mathcal{S})^2}{2\sigma}(k + 1) \qquad \textit{for any } \mathcal{S}: T - |\mathcal{S}| \leq k \textit{ and } \boldsymbol{u} \in \mathcal{W}, \tag{7}$$

where $\widetilde{G}(\boldsymbol{u}, \mathcal{S}) = 2G(\mathcal{S}) + \max_{t:\|\boldsymbol{g}_t\|_2 \leq 2G(\mathcal{S})} \|\nabla f_t(\boldsymbol{u})\|_2$.

The standard regret bound of OGD for strongly convex losses is of order $\frac{G^2}{\sigma}\log T$ (Hazan et al., 2007), so in this case we can tolerate $k = O(\log T)$ outliers without suffering in the rate, under the additional assumption that $\widetilde{G}(\boldsymbol{u}, \mathcal{S}) = O(G(\mathcal{S}))$. This seems like a reasonable assumption if we think of the condition $\|\boldsymbol{g}_t\|_2 \leq 2G(\mathcal{S})$ as expressing that round t is not too extreme.

Huber ϵ-Contamination   As a final example, we consider the Huber ϵ-contamination setting (Huber, 1964). In this case losses are of the form $f_t(\boldsymbol{w}) = f(\boldsymbol{w}, \xi_t)$, where the random variables ξ_t are sampled i.i.d. from a mixture distribution P_ϵ defined by

$$\xi \sim \begin{cases} P & \text{if } M = 0 \\ Q & \text{if } M = 1 \end{cases} \qquad \text{where} \qquad M \sim \text{Bernoulli}(\epsilon)$$

for some $\epsilon \in [0, 1)$. The interpretation is that P is the actual distribution of interest, which is contaminated by outliers drawn from Q. The hidden variable M is not observed by the learner, so it is not known which observations are outliers. Let $\mathcal{S}^* \subseteq [T]$ denote the set of inlier rounds in which $M_t = 0$. Then the robust regret $R_T(\boldsymbol{u}, \mathcal{S}^*)$ may be viewed as the ordinary regret on a modified loss function $\tilde{f}(\boldsymbol{w}, M, \xi)$ that is equal to $f(\boldsymbol{w}, \xi)$ on samples from P but zero on samples from Q, i.e. $\tilde{f}(\boldsymbol{w}, M, \xi) = \mathbf{1}\{M = 0\}f(\boldsymbol{w}, \xi)$ and

$$R_T(\boldsymbol{u}, \mathcal{S}^*) = \sum_{t=1}^{T} \left(\tilde{f}(\boldsymbol{w}_t, M_t, \xi_t) - \tilde{f}(\boldsymbol{u}, M_t, \xi_t) \right). \tag{8}$$

Let the risk with respect to the inlier distribution P be defined as

$$\text{Risk}_P(\boldsymbol{w}) = \underset{\xi \sim P}{\mathbb{E}}[f(\boldsymbol{w}, \xi)].$$

Then, applying online-to-batch conversion (Cesa-Bianchi et al., 2004) to the modified loss $\tilde{f}$, we obtain the following result, which bounds the excess risk under P by the robust regret when the observations are drawn from the contaminated mixture P_ϵ, without requiring any assumptions about the outliers coming from Q:

Lemma 4 (Huber ϵ-Contamination)   *Suppose the losses f_t are i.i.d. according to the mixture distribution P_ϵ, and let $\boldsymbol{u}_P \in \arg\min_{\boldsymbol{w} \in \mathcal{W}} \text{Risk}_P(\boldsymbol{w})$ be the optimal parameters for the distribution of the inliers. Let $\bar{\boldsymbol{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t$, where $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T$ are the predictions of the learner. Then*

$$\underset{P_\epsilon}{\mathbb{E}}\left[\text{Risk}_P(\bar{\boldsymbol{w}}_T) - \text{Risk}_P(\boldsymbol{u}_P)\right] \leq \frac{\mathbb{E}_{P_\epsilon}\left[R_T(\boldsymbol{u}_P, \mathcal{S}^*)\right]}{(1 - \epsilon)T}. \tag{9}$$

Moreover, if $|f(\boldsymbol{w}, \xi) - f(\boldsymbol{u}_P, \xi)| \leq B$ almost surely when $\xi \sim P$ is an inlier, then for any $0 < \delta \leq 1$

$$\text{Risk}_P(\bar{\boldsymbol{w}}_T) - \text{Risk}_P(\boldsymbol{u}_P) \leq \frac{R_T(\boldsymbol{u}_P, \mathcal{S}^*)}{(1 - \epsilon)T} + \frac{2B}{1 - \epsilon}\sqrt{\frac{2}{T}\ln\frac{1}{\delta}} \tag{10}$$

with P_ϵ-probability at least $1 - \delta$.

(Details of the proof are given in Appendix A.) We see that, if we can control the robust regret with respect to the unknown set $\mathcal{S}^*$ of inlier rounds, then we can also control the excess risk with respect to the inlier distribution P. For example, instantiating the learner as in Corollary 2 leads to the following specialization of Lemma 4.

Corollary 5   *In the setting of Lemma 4, suppose that $\|\nabla f(\boldsymbol{w}, \xi)\| \leq G$ for all $\boldsymbol{w} \in \mathcal{W}$ almost surely when $\xi \sim P$ is an inlier, and that $\epsilon \leq 1/2$. Let the learner be instantiated as in Corollary 2 with $k = \lceil \epsilon T + \sqrt{2T\epsilon(1 - \epsilon)\ln(2/\delta)} + \frac{1}{3}(1 - \epsilon)\ln(2/\delta) \rceil$ for any $0 < \delta \leq 1$. Then*

$$\text{Risk}_P(\bar{\boldsymbol{w}}_T) - \text{Risk}_P(\boldsymbol{u}_P) \leq 12DG\epsilon + \frac{2DG\left(5\sqrt{2\ln(2/\delta)} + 2\right)}{\sqrt{T}} + \frac{2DG\left(\ln(2/\delta) + 10\right)}{T} \tag{11}$$

with P_ϵ-probability at least $1 - \delta$.

Here $\nabla f(\boldsymbol{w}, \xi)$ should be read as the gradient of $f(\boldsymbol{w}, \xi)$ with respect to $\boldsymbol{w}$. The constant dependence on $DG\epsilon$, which does not go to zero with increasing T, is unavoidable because P is non-identifiable based on samples from P_ϵ. For instance, consider the linear loss $f(w, \xi) = \xi w$ with $\mathcal{W} = [-D/2, +D/2]$ and P_ϵ such that $\xi = -G$ and $\xi = +G$ both with probability ϵ, and $\xi = 0$ with probability $1 - 2\epsilon$. Then we cannot distinguish the case that $P = P_\epsilon(\cdot \mid \xi \leq 0)$ and Q is a point-mass on $+G$ from the case that $P = P_\epsilon(\cdot \mid \xi \geq 0)$ with Q a point-mass on $-G$. No matter what the output of the learner is, its excess risk under P will always be at least $DG\epsilon$ in one of these two cases.

The proof of Corollary 5 is postponed to Appendix A. It is a straightforward combination of Lemma 4 and Corollary 2, with the only point of attention being the tuning of the number of outliers k. In expectation, the number of outliers is ϵT, but we choose k slightly larger so that the probability that the number of outliers exceeds k is negligible.

3.2. Lower Bounds

We now show that the bounds obtained in the previous part of this section are non-improvable in general. First note that one can always choose $\mathcal{S} = [T]$ (no outliers) and apply a standard lower bound for online learning algorithms which guarantees expected regret $\Omega(\sqrt{T})$ for general losses and $\Omega(\ln T)$ for strongly-convex losses. This matches the first term in the bound of Theorem 1. Therefore, we will only show a bound $\Omega(k)$, which, combined with the standard one, leads to a $\Omega(\max\{\sqrt{T}, k\}) = \Omega(\sqrt{T} + k)$ lower bound on the regret for general convex losses and $\Omega(\ln T + k)$ for strongly convex losses.

Consider a learning task over domain $\mathcal{W} = [-W, W]$ for some $W > 0$. To prove a lower bound for general convex losses, we choose the loss sequence to be $f_t(w) = G\xi_t w$, where $\xi_t \in \{-1, +1\}$ are i.i.d. Rademacher random variables with $\Pr(\xi_t = -1) = \Pr(\xi_t = +1) = \frac{1}{2}$, while $G > 0$ controls the size of the gradients/losses.

Theorem 6 (Lower Bound with I.I.D. Losses) *For any k and any online learning algorithm run on the sequence defined above, there exist adversarial choices of $\mathcal{S}$ with $T - |\mathcal{S}| \leq k$ and $u \in \mathcal{W}$ such that*

$$\underset{f_1, \ldots, f_T}{\mathbb{E}} [R_T(u, \mathcal{S})] \geq \frac{DG(\mathcal{S})k}{4},$$

where $f_1, \ldots, f_T$ are i.i.d. as described above.

Proof Let $S_1 = \{t \in [k] : \xi_t = 1\}$ and $S_{-1} = \{t \in [k] : \xi_t = -1\}$. The adversary will choose $u = -W\zeta$ and $\mathcal{S} = S_\zeta \cup \{k+1, \ldots, T\}$, where $\zeta \in \{-1, 1\}$ is a Rademacher random variable independent of $\xi_1, \ldots, \xi_T$. The expected regret jointly over $\xi_1, \ldots, \xi_T, \zeta$ is then given by

$$\mathbb{E}[R_T(u, \mathcal{S})] = \mathbb{E}\left[G\sum_{t=1}^{T} \mathbf{1}_{t \in \mathcal{S}} w_t \xi_t - G\sum_{t=1}^{T} \mathbf{1}_{t \in \mathcal{S}} u\xi_t\right]$$

$$= G\sum_{t=1}^{k} \underbrace{\mathbb{E}[\mathbf{1}_{\zeta = \xi_t} w_t \xi_t]}_{=0} + G\sum_{t=k+1}^{T} \underbrace{\mathbb{E}[w_t \xi_t]}_{=0} + GW\sum_{t=1}^{k} \underbrace{\mathbb{E}[\mathbf{1}_{\zeta = \xi_t} \zeta \xi_t]}_{=1/2} + GW\sum_{t=k+1}^{T} \underbrace{\mathbb{E}[\zeta \xi_t]}_{=0}$$

$$= GW\frac{k}{2},$$

9

where we used the independence of ξ_t and ζ in the second and the fourth sum, while

$$\mathbb{E}\left[\mathbf{1}_{\zeta=\xi_t} w_t \xi_t\right] = \mathbb{E}\left[\,\mathbb{E}\left[\mathbf{1}_{\zeta=\xi_t} w_t \xi_t \,|\, \xi_t\right]\,\right] = \mathbb{E}\left[w_t \xi_t/2\right] = 0, \quad \text{and } \mathbb{E}\left[\mathbf{1}_{\zeta=\xi_t} \zeta \xi_t\right] = \mathbb{E}\left[\mathbf{1}_{\zeta=\xi_t}\right] = \frac{1}{2}.$$

As the bound holds for the random choice of ζ it also holds for the worst-case choice of ζ. The theorem now follows from $D = \max_{u,w\in\mathcal{W}} |w-u| = 2W$ and $G(\mathcal{S}) = \max_{t\in\mathcal{S}} |g_t| = \max_{t\in\mathcal{S}} G|\xi_t| = G$. ∎

A similar bounding technique leads to a lower bound for σ-strongly convex losses, except that the distribution of the losses differs between the first k rounds and the later rounds. This still implies a lower bound for adversarially generated data, but not for i.i.d. losses. In this case, we will choose the domain $\mathcal{W} = [-W, W]$, the loss sequence based on the σ-strongly convex squared loss, $f_t(w) = \frac{\sigma}{2}(w - W\xi_t)^2$, for $t \le k$, and $f_t(w) = \frac{\sigma}{2}(w - W\zeta)^2$ for $t \ge k$, where $\xi_1, \dots \xi_k$ and ζ are again i.i.d. Rademacher variables.

Theorem 7 (Lower Bound for Strongly Convex Losses) *For any k and any online learning algorithm, there exist adversarial choices of $\mathcal{S}$ with $T - |\mathcal{S}| \le k$ and $u \in \mathcal{W}$ such that*

$$\underset{f_1,\dots,f_T}{\mathbb{E}} [R_T(u,\mathcal{S})] \ge \frac{G^2(\mathcal{S})k}{16\sigma},$$

where $f_1, \dots, f_T$ are the σ-strongly convex losses described above.

Proof Using the same notation as in the proof of Theorem 6, the adversary will choose $u = W\zeta$ and $\mathcal{S} = S_\zeta \cup \{t+1, \dots, T\}$. The expected regret jointly over $\xi_1, \dots, \xi_k, \zeta$ is given by

$$\mathbb{E}\left[R_T(u,\mathcal{S})\right] = \frac{\sigma}{2}\sum_{t=1}^{k}\underbrace{\mathbb{E}\left[\mathbf{1}_{\zeta=\xi_t}(w_t - W\xi_t)^2\right]}_{\ge W^2/2} + \frac{\sigma}{2}\sum_{t=k+1}^{T}\underbrace{\mathbb{E}\left[(w_t - W\zeta)^2\right]}_{\ge 0}$$

$$- \frac{\sigma}{2}\sum_{t=1}^{k}\underbrace{\mathbb{E}\left[\mathbf{1}_{\zeta=\xi_t}(\zeta - W\xi_t)^2\right]}_{=0} \ge \frac{\sigma W^2 k}{4},$$

where to bound the first sum we used

$$\mathbb{E}\left[\mathbf{1}_{\zeta=\xi_t}(w_t - W\xi_t)^2\right] = \mathbb{E}\left[\,\mathbb{E}\left[\mathbf{1}_{\zeta=\xi_t}(w_t - W\xi_t)^2 \,|\, \xi_t\right]\,\right] = \mathbb{E}\left[(w_t - W\xi_t)^2/2\right]$$
$$= \mathbb{E}\left[w_t^2/2 - W\xi_t w_t + W^2/2\right] = w_t^2/2 + W^2/2 \ge W^2/2.$$

To finish the proof note that $|\nabla f_t(w_t)| = \sigma|w_t - W\xi_i| \le 2\sigma W$ so that $G(\mathcal{S}) \le 2\sigma W$. ∎

4. Robustness for Quantiles

In this section we consider robust online linear optimization in the stochastic i.i.d. setting. That is, we consider i.i.d. gradients $g_t \sim \mathbb{P}$ that are in particular independent of the learner's prediction w_t. Let $G_p := q_p(\|g\|_*)$ be the p-quantile of the gradient in dual norm $\|\cdot\|_*$. To keep things simple, we will assume that $\mathbb{P}$ does not have an atom at G_p, so that $\mathbb{P}\{\|g\|_* \le G_p\} = p$ exactly. We call

a gradient g_t an *outlier* if $\|g_t\|_* > G_p$. Fix a domain $\mathcal{W}$ of diameter D in the norm $\|\cdot\|$. We are interested in algorithms that know $\mathcal{W}$ and p but not G_p, play $w_t \in \mathcal{W}$, and we aim to bound their expected robust regret on the (random!) set of inliers $\mathcal{S} = \{t \in [T] : \|g_t\|_* \leq G_p\}$. That is, we aim to control

$$\bar{R}_T := \mathbb{E}\left[\max_{u \in \mathcal{W}} R_T(u, \mathcal{S})\right] = \mathbb{E}\left[\max_{u \in \mathcal{W}} \sum_{t \in [T]:\|g_t\|_* \leq G_p} \langle w_t - u, g_t \rangle\right]. \tag{12}$$

Note that a bound on the expected robust regret implies a robust pseudo-regret bound, where the data-dependent maximum is replaced by the fixed minimiser of the expected loss on inliers, i.e. $u^* \in \arg\min_{u \in \mathcal{W}} u^\mathsf{T} \mathbb{E}[g_t \mid \|g_t\|_* \leq G_p]$. Our FILTER algorithm for the stochastic setting is shown as Algorithm 2. The main idea is that it only passes rounds to the base ALG for which it is virtually certain that they are inliers. To this end our FILTER computes a lower confidence bound LCB_t on the quantile G_p. Smaller gradients are included, while larger ones are discarded. The crux of the robust regret bound proof is then dealing with the inlier gradients that end up being dropped. We will find it instructive to state our algorithms and confidence bounds with a free confidence parameter δ. Tuning our approach will then lead us to set $\delta = T^{-2}$.

Algorithm 2: Filtering meta algorithm for Robust Quantile Regret

---

Input: Quantile level $p \in (0, 1)$, confidence δ, online learner ALG
for $t = 1, 2, \ldots$ **do**

    Have ALG produce w_t. Receive gradient g_t;

    Let $\hat{q}_{t-1}$ be the empirical quantile function of past gradients $g_1, \ldots, g_{t-1}$. ;

    Compute $\text{LCB}_{t-1} = \hat{q}_{t-1}(p - u_{t-1})$ at threshold $u_{t-1} = \sqrt{t^{-1}2p(1-p)\ln\frac{1}{\delta}} + \frac{1}{3}t^{-1}\ln\frac{1}{\delta}$;

    if $\|g_t\|_* \leq \text{LCB}_{t-1}$ **then**

        Pass round t on to ALG;

    else

        Ignore round t;

    end

end

We now show that the expected robust regret is small.

Theorem 8 *Let ALG have individual sequence regret bound $B_T(G)$ for T rounds with gradients of dual norm at most G, and which is concave in T. Let D be the diameter of the domain. Then the FILTER Meta-Algorithm 2 with $\delta = T^{-2}$ has expected robust regret bounded by*

$$\bar{R}_T \leq B_{pT}(G_p) + DG_p\left(4\sqrt{2p(1-p)T\ln T} + \frac{13}{3}(\ln T)^2 + 3\right).$$

If ALG does its job, the first term is the minimax optimal regret for when the outlier rounds were known. The other terms quantify the cost of being robust. When p is not extreme, this cost is of order $G_p D\sqrt{T\ln T}$, rendering it the dominant term overall (escalating the minimax regret by a mild log factor). When p tends to 1 or 0, the robustness overhead gracefully reduces to the $(\ln T)^2$ regime.

The proof can be found in Appendix B. The main ideas are as follows. As we have no control over outlier gradients (they may be astronomical), we must assume that ALG gets confused without recourse if FILTER ever passes it any outlier. Note that FILTER is not evaluated on outlier rounds, so it does not suffer from this gradient's *magnitude*. But its effect is that, for all we know, ALG is rendered forever useless, upon which FILTER may incur the maximum possible regret of $G_p D T$. Our approach will be to choose our threshold for inclusion conservatively, and to apply concentration in all rounds simultaneously, to ensure this bad event is rare (this is the source of the $\ln T$ factor). A second concentration allows us to deal with the discarded inliers.

We conclude the section with a selection of remarks.

Examples The examples of Section 3.1.1 also apply here. Depending on the setting, and hence the appropriate base algorithm ALG, the dominant regret term can be either the $D G_p \sqrt{p(1-p) T \ln T}$ term, or the $B_{pT}(G_p)$ term. The former case applies for OGD, while the latter case happens in the K-experts setting with many experts and few rounds, i.e. $K \gg T$. There adding robustness comes essentially for free.

Anytime Robust Regret As stated, the algorithm needs to know the horizon T up front to set the confidence parameter δ in the deviation width u_t. We can use a standard doubling trick on T to get an anytime algorithm.

Anytime concentration One may wonder how much the analysis can be improved by replacing our union bound over time steps with a time-uniform Bernstein concentration inequality, as e.g. developed by Howard and Ramdas (2019). Sadly, the best we can hope for is to be able to use $\delta = \frac{1}{T}$, which would lead to a constant factor $\sqrt{2}$ improvement on the dominant term. We cannot tolerate a higher overall failure probability, for we have to pacify the regret upon failure, which may be of order T.

High Probability Version Going into the proof, we see that a high probability robust regret bound is also possible. We would need to change the analysis of $P^{(2)}$, as we currently analyse it in expectation. Observing that it is a sum of T conditionally independent increments, we may use martingale concentration to find that, with probability at least $1 - T^{-1}$, this sum is at most its mean (which features in the expected regret bound) plus a deviation of order $\sqrt{T \ln T}$. We obtain a high-probability analogue of Theorem 8 with slightly inflated constant.

Large-Feature-Vectors-as-Outliers We may also deal with non-i.i.d. gradients using exactly the same techniques developed above, as follows. We assume that $f_t(\boldsymbol{w}) = h_t(\boldsymbol{w}^\intercal \boldsymbol{X}_t)$, where $\boldsymbol{X}_t \in \mathbb{R}^d$ is a feature vector available at the beginning of round t, and h_t is a scalar Lipschitz convex loss function, revealed at the end of round t. This setting includes e.g. linear classification with hinge or logistic loss. Upon assuming that feature vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ are drawn i.i.d. from $\mathbb{P}$ (while the h_t are arbitrary, possibly adversarially chosen), we can take the p-quantile $X_p \coloneqq q_p(\|\boldsymbol{X}\|_*)$ of the dual norm of the feature vectors. We may then measure the robust expected regret (12) on the inlier rounds $\mathcal{S} = \{t \in [T] : \|\boldsymbol{X}_t\|_* \le X_p\}$, and obtain the analogue of Theorem 8, where the only subtlety is using the gradient bound on h_t to transfer from inlier $\boldsymbol{X}_t$ to small loss.

Proposition 9 *Consider a joint distribution on sequences of feature vectors and scalar Lipschitz convex functions $(\boldsymbol{X}_1, h_1), (\boldsymbol{X}_2, h_2), \ldots$ such that the feature vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ are i.i.d. with distribution $\mathbb{P}$ on $\mathbb{R}^d$.[1] Let $X_p = q_p(\|\boldsymbol{X}\|_*)$ be the p-quantile of the feature dual norm. Let ALG be an algorithm for online-convex optimisation over a domain of diameter D and loss functions*

---

1. We do not constrain the distribution of $h_1, h_2, \ldots$, so we can model adversarial loss functions that are correlated with the feature vectors.

$f_t(\boldsymbol{w}) = h_t(\boldsymbol{w}^\intercal \boldsymbol{X}_t)$ *that guarantees individual-sequence regret bounded by* $B_T(X)$ *in any* T-*round interaction with* $\|\boldsymbol{X}_t\|_* \leq X$, *without having to know* X *up front. Consider FILTER Meta-Algorithm* 2 *with* $\boldsymbol{g}_t$ *replaced by* $\boldsymbol{X}_t$. *Then the expected robust regret on inlier rounds* $\mathcal{S} = \{t \in [T] : \|\boldsymbol{X}_t\|_* \leq X_p\}$ *is bounded by*

$$\bar{R}_T = \mathbb{E}\left[\max_{\boldsymbol{u} \in \mathcal{W}} \sum_{t \in \mathcal{S}} (f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u}))\right] \leq B_{pT}(X_p) + DX_p\left(4\sqrt{2p(1-p)T \ln T} + \frac{13}{3}(\ln T)^2 + 3\right).$$

Proof The proof follows that of Theorem 8, with one extra (standard) step. Namely, to bound the loss on inlier rounds (for the dropped rounds term $P^{(2)}$ in the proof, and the concentration failure term $P^{(3)}$ in the proof), we use convexity, Hölder and bounded derivative to obtain

$$f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u}^*) \leq h_t(\boldsymbol{w}_t^\intercal \boldsymbol{X}_t) - h_t(\boldsymbol{u}^{*\intercal} \boldsymbol{X}_t) \leq h_t'(\boldsymbol{w}_t^\intercal \boldsymbol{X}_t)(\boldsymbol{w}_t - \boldsymbol{u}^*)^\intercal \boldsymbol{X}_t \leq DX_p.$$

∎

Online-to-Batch Example We now discuss an example where the standard theory for stochastic gradient descent does not apply, but the iterate average of online gradient descent with quantile-based filtering still gives risk convergence guarantees. To keep things simple, we work in the one-dimensional setting with $\mathcal{W} = [-1, +1]$. To stay within the assumptions of Proposition 9, we take f_t to be the logistic loss $f_t(w) = h_t(wX_t)$ with $h_t(z) = \ln(1 + e^{-y_t z})$ for $y_t \in \{-1, +1\}$. To make things interesting, we take $\boldsymbol{X}_t \in \mathbb{R}$ to have a distribution with heavy tails, with $\mathbb{P}(|\boldsymbol{X}_t| > x)$ of order $x^{-(1+\gamma)}$ for large enough x, for some $\gamma \in (0, 1)$. Taking $\gamma > 0$ ensures that the expected loss $\mathbb{E}[f_t(w)]$ is finite (as $f_t(w) \approx (-\boldsymbol{X}_t y_t w)_+$ for large $\boldsymbol{X}_t$), and hence has a bonafide minimiser (which can be in the interior or on the boundary, depending on the details of the distribution). Taking $\gamma < 1$ ensures that the tails are so heavy that $\mathbb{E}[f_t'(w)^2] = \infty$ (as $f_t'(w) \approx (-y_t \boldsymbol{X}_t)_+$ for large $\boldsymbol{X}_t$), and hence standard theory for SGD does not apply. Instead we will use Lemma 4 and Proposition 9 to argue that the filtered iterate average $\bar{w}_T$ approximates the minimiser of the risk u^* in the sense that

$$\mathbb{E}_\mathbb{P}\left[\text{Risk}_\mathbb{P}(\bar{w}_T) - \text{Risk}_\mathbb{P}(u^*)\right] \to 0 \quad \text{as} \quad T \to \infty. \tag{13}$$

To bound the risks above, we will decompose $\mathbb{P} = pP + (1-p)Q$ where p is a quantile level chosen below, $P = \mathbb{P}(\cdot \big| \|\boldsymbol{X}_t\| \leq X_p)$ and $Q = \mathbb{P}(\cdot \big| \|\boldsymbol{X}_t\| > X_p)$. We will bound the $\mathbb{P}$-risks in terms of P-risks, then we will use Lemma 4 to bound the P-risk difference in terms of the robust regret, and we will use Proposition 9 to bound that regret. We will settle on picking $p = 1 - \frac{1}{\sqrt{T}}$. This has the effect that the p-quantile is $X_p \propto T^{\frac{1}{2(1+\gamma)}}$ (by inverting the tail probability). On the one hand, for any $w \in \mathcal{W}$, the bias, i.e. the difference in risk on P (inliers only) and on $\mathbb{P}$ (full distribution), is at most of order

$$|\text{Risk}_\mathbb{P}(w) - p \, \text{Risk}_P(w)| \leq \mathbb{E}_\mathbb{P}\left[f_t(w)\mathbf{1}_{|\boldsymbol{X}_t|>X_p}\right] \approx \mathbb{E}_\mathbb{P}\left[(-\boldsymbol{X}_t y_t w)\mathbf{1}_{|\boldsymbol{X}_t|>X_p}\right]$$

$$\leq \mathbb{E}_\mathbb{P}\left[|\boldsymbol{X}_t|\mathbf{1}_{|\boldsymbol{X}_t|>X_p}\right] = \int_{X_p}^\infty \mathbb{P}(|\boldsymbol{X}_t| > x)\,\mathrm{d}x \propto X_p^{-\gamma} \propto T^{-\frac{\gamma}{2(1+\gamma)}}.$$

On the other hand, the regret bound for T-round online gradient descent with gradient norms bounded by X is $B_T(X) = O(DX\sqrt{T})$. Hence for our choice of p, the first term in the bound from Proposition 9 is dominant and of order $X_p\sqrt{T}$. Dividing by T to plug in to Lemma 4 results in $\frac{X_p\sqrt{T}}{T} \propto T^{-\frac{\gamma}{2(1+\gamma)}}$. Both contributions (bias and regret) are of the same order and converge to zero, indicating that quantile-filtered online gradient descent achieves (13).

5. Conclusion and Future Work

We have shown that the robust regret can be controlled for adversarial data when there are at most k outliers. A general question that we leave open is whether it is possible to get a bound for adversarial losses that does not depend on the number of outliers k, but on some other natural property of the losses. For instance, we may try to incorporate prior knowledge about the size of the gradients by specifying a prior π on gradient norms and bounding the robust regret in terms of the prior probability $\pi(G(\mathcal{S}))$ of the size of the inlier gradients. A possible way to approach this might be to introduce specialist experts for different thresholds G and then aggregate these. This runs into severe difficulties, however, because we only find out whether a specialist should be active or not in round t *after* making our prediction $\boldsymbol{w}_t$ and observing $\boldsymbol{g}_t$. Moreover, specialists would have different loss ranges and the robust regret can only depend on the loss range $G(\mathcal{S})$ of the correct specialist.

We also provided a sublinear bound on the robust regret for i.i.d. gradients when the outliers are defined as rounds in which the gradients exceed their p-quantile, or when they can be bounded in terms of an i.i.d. variable $\boldsymbol{X}_t$. Alternatively, outliers might be defined as gradients with norms exceeding their empirical p-quantile at the end of T rounds. For i.i.d. gradients, the empirical p-quantile after T rounds is close to the actual p-quantile with high probability, so this case can be handled by running the method from Section 4 for a slightly inflated p. However, the empirical quantile formulation continues to make sense even when gradients are not i.i.d., so it would be interesting to know whether a linear number of outliers can be tolerated in any such non-i.i.d. cases.

Acknowledgments

References

Idan Amir, Idan Attias, Tomer Koren, Yishay Mansour, and Roi Livni. Prediction with corrupted expert advice. In *Advances in Neural Information Processing Systems 33*, 2020.

Pranjal Awasthi, Maria Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63(6), 2017.

Peter L. Bartlett, Elad Hazan, and Alexander Rakhlin. Adaptive online gradient descent. In *NeurIPS 20*, pages 65–72, 2007.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3), 2011.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48:1148–1185, 2012.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 47–60. ACM, 2017.

Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Online and distribution-free robustness: Regression and contextual bandits with Huber contamination. *CoRR*, abs/2010.04157, 2020. URL https://arxiv.org/abs/2010.04157.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, pages 2755–2771. Society for Industrial and Applied Mathematics, 2019.

Alexey Chernov and Vladimir Vovk. Prediction with expert evaluators' advice. In *Proceedings of the 20th international conference on Algorithmic learning theory*, ALT '09, pages 8–22, Berlin, Heidelberg, 2009. Springer-Verlag.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1405–1411, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/daniely15.html.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(37):1281–1316, 2014.

Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019. URL http://arxiv.org/abs/1911.05911.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a Gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 2683–2702, USA, 2018. Society for Industrial and Applied Mathematics.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1596–1606. PMLR, 2019a.

Ilias Diakonikolas, Weihao Kong, , and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019b.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth. Using and combining predictors that specialize. In *Proc. 29th Annual ACM Symposium on Theory of Computing*, pages 334–343. ACM, 1997.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1562–1578. PMLR, 2019.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325, 2015.

Elad Hazan and C. Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 393–400. ACM, 2009.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Steven R. Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to a/b-testing and best-arm identification, 2019.

Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 1964.

Peter J. Huber. *Robust statistics*. Wiley, 1981.

Adam Klivans, Pravesh K. Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1420–1430. PMLR, 2018.

Adam R. Klivans, Philip M. Long, and Rocco A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(94):2715–2740, 2009. URL http://jmlr.org/papers/v10/klivans09a.html.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJGU3Rodl.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674, 2016.

Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 411–421. PMLR, 2020.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(3):447–461, 2016.

G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Found Comput Math 19*, 19:1145–1190, 2019a. doi: https://doi.org/10.1007/s10208-019-09427-x.

Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.*, 47(2):783–794, 2019b.

Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean, 2019c.

Thodoris Lykouris, Vahab S. Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 114–122, 2018.

Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Alexander V. Nazin, Arkadi S. Nemirovsky, B. Tsybakov, Alexandre, and Anatoli B. Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80:1607–1627, 2019.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018. Special Issue on ALT 2015.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *CoRR*, abs/1802.06485, 2018. URL http://arxiv.org/abs/1802.06485.

Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A robust univariate mean estimator is all you need. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4034–4044. PMLR, 2020.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

John W. Tukey. A survey of sampling from contaminated distributions. STRG Technical report 33, Princeton University, 1959.

Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Yizhen Wang and Kamalika Chaudhuri. Data poisoning attacks against online learning, 2018.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings*

of Machine Learning Research, pages 6677–6686. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/ward19a.html.

Xuezhou Zhang, Xiaojin Zhu, and Laurent Lessard. Online data poisoning attacks. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 201–210. PMLR, 2020.

Appendix A. Proofs for Examples from Section 3

Proof of Corollary 2 Let $\mathcal{P} \subset [T]$ denote the rounds that are passed on to OGD. Then the linearized regret of OGD on the rounds in $\mathcal{P}$ is bounded by

$$B_T(2G(\mathcal{S})) \leq 2D\sqrt{\sum_{t \in \mathcal{P}} \|\boldsymbol{g}_t\|_2^2},$$

as follows from arguments similar to those by Duchi et al. (2011) (see e.g. Corollary 2 by Orabona and Pál (2018)). Bounding further, we obtain

$$B_T(2G(\mathcal{S})) \leq 2D\sqrt{\sum_{t \in \mathcal{S}} \|\boldsymbol{g}_t\|_2^2 + \sum_{t \in \mathcal{P} \setminus \mathcal{S}} \|\boldsymbol{g}_t\|_2^2} \leq 2D\sqrt{\sum_{t \in \mathcal{S}} \|\boldsymbol{g}_t\|_2^2} + 2D\sqrt{\sum_{t \in \mathcal{P} \setminus \mathcal{S}} \|\boldsymbol{g}_t\|_2^2}$$

$$\leq 2D\sqrt{\sum_{t \in \mathcal{S}} \|\boldsymbol{g}_t\|_2^2} + 2DG(\mathcal{S})\sqrt{k},$$

where the last step uses that $|\mathcal{P} \setminus \mathcal{S}| \leq |[T] \setminus \mathcal{S}| \leq k$ by assumption on $\mathcal{S}$. Plugging this into (4) and bounding $R_T(\boldsymbol{u}, \mathcal{S}) \leq \widetilde{R}_T(\boldsymbol{u}, \mathcal{S})$ and $D(\boldsymbol{u}, \mathcal{S}) \leq D$, the first inequality in (6) follows. Finally, using that $\|\boldsymbol{g}_t\|_2 \leq G(\mathcal{S})$ for all $t \in \mathcal{S}$, we see that the second inequality holds as well. ∎

Proof of Corollary 3 Let $\mathcal{P} \subset [T]$ denote the rounds that are passed on to ALG. By the proof of Theorems 2.1 and 4.1 of Bartlett et al. (2007), the linearized regret of ALG on the rounds in $\mathcal{P}$ is bounded by

$$\widetilde{R}_T(\boldsymbol{u}, \mathcal{P}) \leq \frac{1}{2} \sum_{t \in \mathcal{P}} \frac{\|\boldsymbol{g}_t\|_2^2}{\sigma t} + \frac{\sigma}{2} \sum_{t \in \mathcal{P}} \|\boldsymbol{w}_t - \boldsymbol{u}\|_2^2 \leq \frac{2G(\mathcal{S})^2}{\sigma}\big(\ln T + 1\big) + \frac{\sigma}{2} \sum_{t \in \mathcal{P}} \|\boldsymbol{w}_t - \boldsymbol{u}\|_2^2.$$

Plugging this into Theorem 1 and applying the definition of strong convexity, we get that the robust regret is bounded by

$$R_T(\boldsymbol{u}, \mathcal{S}) \leq \widetilde{R}_T(\boldsymbol{u}, \mathcal{S}) - \frac{\sigma}{2} \sum_{t \in \mathcal{S}} \|\boldsymbol{w}_t - \boldsymbol{u}\|_2^2$$

$$\leq \frac{2G(\mathcal{S})^2}{\sigma}\big(\ln T + 1\big) + 4D(\boldsymbol{u}, \mathcal{S})G(\mathcal{S})(k+1) + \frac{\sigma}{2} \sum_{t \in \mathcal{P}} \|\boldsymbol{w}_t - \boldsymbol{u}\|_2^2 - \frac{\sigma}{2} \sum_{t \in \mathcal{S}} \|\boldsymbol{w}_t - \boldsymbol{u}\|_2^2$$

$$\leq \frac{2G(\mathcal{S})^2}{\sigma}\big(\ln T + 1\big) + 4D(\boldsymbol{u}, \mathcal{S})G(\mathcal{S})(k+1) + \frac{\sigma}{2} \sum_{t \in \mathcal{P} \setminus \mathcal{S}} \|\boldsymbol{w}_t - \boldsymbol{u}\|_2^2$$

$$\leq \frac{2G(\mathcal{S})^2}{\sigma}\big(\ln T + 1\big) + 4D(\boldsymbol{u}, \mathcal{S})G(\mathcal{S})(k+1) + \frac{\sigma D(\boldsymbol{u}, \mathcal{S})^2}{2}k.$$

From this the desired result follows because $G(\mathcal{S}) \leq \widetilde{G}(\boldsymbol{u}, \mathcal{S})/2$ and

$$D(\boldsymbol{u}, \mathcal{S}) \leq \max_{t:\|\boldsymbol{g}_t\|_2 \leq 2G(\mathcal{S})} \frac{\|\boldsymbol{g}_t\|_2 + \|\nabla f_t(\boldsymbol{u})\|_2}{\sigma} \leq \frac{\widetilde{G}(\boldsymbol{u}, \mathcal{S})}{\sigma}$$

by Lemma 10 below. ∎

Lemma 10 *Suppose f_t is σ-strongly convex. Then $\|\boldsymbol{w} - \boldsymbol{u}\|_2 \leq \frac{\|\nabla f_t(\boldsymbol{w})\|_2 + \|\nabla f_t(\boldsymbol{u})\|_2}{\sigma}$ for all $\boldsymbol{w}, \boldsymbol{u} \in \mathcal{W}$.*

Proof Applying the definition of σ-strong convexity twice, we have

$$(\boldsymbol{w} - \boldsymbol{u})^\mathsf{T} \nabla f_t(\boldsymbol{u}) + \frac{\sigma}{2}\|\boldsymbol{w} - \boldsymbol{u}\|_2^2 \leq f_t(\boldsymbol{w}) - f_t(\boldsymbol{u}) \leq (\boldsymbol{w} - \boldsymbol{u})^\mathsf{T} \nabla f_t(\boldsymbol{w}) - \frac{\sigma}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2,$$

which leads to

$$\begin{aligned}
\sigma\|\boldsymbol{w} - \boldsymbol{u}\|_2^2 &\leq (\boldsymbol{w} - \boldsymbol{u})^\mathsf{T}\big(\nabla f_t(\boldsymbol{w}) - \nabla f_t(\boldsymbol{u})\big) \\
&\leq \|\boldsymbol{w} - \boldsymbol{u}\|_2 \|\nabla f_t(\boldsymbol{w}) - \nabla f_t(\boldsymbol{u})\|_2 \leq \|\boldsymbol{w} - \boldsymbol{u}\|_2\big(\|\nabla f_t(\boldsymbol{w})\|_2 + \|\nabla f_t(\boldsymbol{u})\|_2\big),
\end{aligned}$$

from which the result follows. ∎

Proof of Lemma 4 Let

$$\widetilde{\mathrm{Risk}}_{P_\epsilon}(\boldsymbol{w}) = \mathbb{E}_{(M,\xi)\sim P_\epsilon}\left[\tilde{f}(\boldsymbol{w}, M, \xi)\right]$$

denote the risk for the modified loss under the mixture distribution P_ϵ. Then the key to both results is to observe that

$$\mathrm{Risk}_P(\bar{\boldsymbol{w}}_T) - \mathrm{Risk}_P(\boldsymbol{u}_P) = \frac{\widetilde{\mathrm{Risk}}_{P_\epsilon}(\bar{\boldsymbol{w}}_T) - \widetilde{\mathrm{Risk}}_{P_\epsilon}(\boldsymbol{u}_P)}{1 - \epsilon}.$$

This allows us to apply standard results for online-to-batch conversion to the modified losses $\tilde{f}$ under distribution P_ϵ: the first inequality follows by combining

$$\mathbb{E}_{P_\epsilon}\left[\widetilde{\mathrm{Risk}}_{P_\epsilon}(\bar{\boldsymbol{w}}_T) - \widetilde{\mathrm{Risk}}_{P_\epsilon}(\boldsymbol{u}_P)\right] \leq \frac{\mathbb{E}_{P_\epsilon}\left[\sum_{t=1}^{T}\left(\tilde{f}(\boldsymbol{w}_t, M_t, \xi_t) - f(\boldsymbol{u}_P, M_t, \xi_t)\right)\right]}{T},$$

with (8), and the second result follows by applying Corollary 2 of Cesa-Bianchi et al. (2004) to the modified excess losses $\tilde{f}(\boldsymbol{w}, M, \xi) - \tilde{f}(\boldsymbol{u}_P, M, \xi)$. By the boundedness assumption on the original excess loss a.s. under P, these are bounded in $[-B, B]$ a.s. under P_ϵ, and we obtain

$$\widetilde{\mathrm{Risk}}_{P_\epsilon}(\bar{\boldsymbol{w}}_T) - \widetilde{\mathrm{Risk}}_{P_\epsilon}(\boldsymbol{u}_P) \leq \frac{\sum_{t=1}^{T}\left(\tilde{f}(\boldsymbol{w}_t, M_t, \xi_t) - \tilde{f}(\boldsymbol{u}_P, M_t, \xi_t)\right)}{T} + 2B\sqrt{\frac{2}{T}\ln\frac{1}{\delta}}$$

with P_ϵ-probability at least $1 - \delta$. The second result then follows by plugging in (8) again. ∎

Proof of Corollary 5 Let $\mathcal{A}$ be the event that (10) holds with $B = DG$ and δ replaced by $\delta/2$, and let $\mathcal{B}$ be the event that the number of outliers $\sum_{t=1}^T M_t$ as at most k. Then Corollary 2 implies that (11) holds on the intersection of $\mathcal{A}$ and $\mathcal{B}$, because

$$
\begin{aligned}
\operatorname{Risk}_P(\bar{\boldsymbol{w}}_T) - \operatorname{Risk}_P(\boldsymbol{u}_P) &\leq \frac{2DG\left(\sqrt{T} + 2k + \sqrt{k} + 2\right)}{(1-\epsilon)T} + \frac{2DG}{1-\epsilon}\sqrt{\frac{2}{T}\ln\frac{2}{\delta}} \\
&\leq \frac{2DG\left(\sqrt{T} + 3k + 2\right)}{(1-\epsilon)T} + \frac{2DG}{1-\epsilon}\sqrt{\frac{2}{T}\ln\frac{2}{\delta}} \\
&\leq \frac{6DG\epsilon}{1-\epsilon} + \frac{2DG\left(1 + 3\sqrt{2\epsilon(1-\epsilon)\ln(2/\delta)}\right)}{(1-\epsilon)\sqrt{T}} + \frac{2DG\left((1-\epsilon)\ln(2/\delta) + 5\right)}{(1-\epsilon)T} + \frac{2DG}{1-\epsilon}\sqrt{\frac{2}{T}\ln\frac{2}{\delta}} \\
&\leq 12DG\epsilon + \frac{2DG\left(2 + 5\sqrt{2\ln(2/\delta)}\right)}{\sqrt{T}} + \frac{2DG\left(\ln(2/\delta) + 10\right)}{T},
\end{aligned}
$$

where the last inequality uses the assumption that $\epsilon \leq 1/2$ to obtain a simpler expression. Now the probability of $\mathcal{A}$ is at least $1 - \delta/2$ by the second result of Lemma 4, which applies because

$$-DG \leq -(\boldsymbol{w} - \boldsymbol{u}_P)^\mathsf{T}\nabla f(\boldsymbol{u}_P, \xi) \leq f(\boldsymbol{w}, \xi) - f(\boldsymbol{u}_P, \xi) \leq (\boldsymbol{w} - \boldsymbol{u}_P)^\mathsf{T}\nabla f(\boldsymbol{w}, \xi) \leq DG \quad (14)$$

P-almost surely. And the probability of $\mathcal{B}$ is at least $1 - \delta/2$ by Bernstein's inequality (Boucheron et al., 2013, Section 2.8). Hence, by the union bound, it follows that both $\mathcal{A}$ and $\mathcal{B}$ hold simultaneously with probability at least $1 - \delta$, as required. ∎

Appendix B. Proof of Theorem 8 from Section 4

Proof The point of departure is that, by definition, $\mathbf{1}\left\{\|\boldsymbol{g}_t\|_* \leq G_p\right\}$ is i.i.d. Bernoulli-p. By Bernstein's concentration inequality (for fixed time t), we have that w.p. $\geq 1 - \delta$

$$\frac{1}{t}\sum_{s=1}^t \mathbf{1}\left\{\|\boldsymbol{g}_s\|_* \leq G_p\right\} \geq p - \sqrt{\frac{2p(1-p)\ln\frac{1}{\delta}}{t}} - \frac{\ln\frac{1}{\delta}}{3t}$$

Rephrasing this event with the empirical quantile function $\hat{q}_t$, we see that w.p. $\geq 1 - \delta$,

$$G_p \geq \operatorname{LCB}_t := \hat{q}_t(p - u_t) \qquad \text{where} \qquad u_t := \sqrt{\frac{2p(1-p)\ln\frac{1}{\delta}}{t}} + \frac{\ln\frac{1}{\delta}}{3t}. \quad (15)$$

We apply the analogous concentration in the other direction to find that with probability at least $1 - \delta$,

$$G_{p - 2u_t - 3/(2t)\ln\frac{1}{\delta}} \leq \operatorname{LCB}_t \quad (16)$$

where the extra margin $\frac{3\ln\frac{1}{\delta}}{2t}$ is necessary to correct for the fact that $p - 2u_t$ may be closer to $1/2$ than p, and hence require a slightly enlarged confidence width. In the remainder of the proof, we fix $\delta = T^{-2}$, to ensure that the probability of failure of either event (15) or (16) over the course of T rounds is at most $\frac{2}{T}$.

Let $\mathcal{E}_t$ denote the event that (15) and (16) hold at round t, and let $\mathcal{E} = \bigcap_{t=1}^{T} \mathcal{E}_t$. We split the expected robust regret in three parts $\bar{R}_T \leq P^{(1)} + P^{(2)} + P^{(3)}$ spelled out below, depending on whether the desired concentration event $\mathcal{E}$ holds or not, and within $\mathcal{E}$ we split the rounds in those where FILTER passes the gradient on to ALG and those were it was ignored:

$$
P^{(1)} := \mathbb{E}\left[\mathbf{1}_{\mathcal{E}} \max_{\boldsymbol{u} \in \mathcal{W}} \sum_{t \in [T] : \|\boldsymbol{g}_t\|_* \leq \mathrm{LCB}_{t-1}} \langle \boldsymbol{w}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle \right]
$$

$$
P^{(2)} := \mathbb{E}\left[\mathbf{1}_{\mathcal{E}} \max_{\boldsymbol{u} \in \mathcal{W}} \sum_{t \in [T] : \mathrm{LCB}_{t-1} < \|\boldsymbol{g}_t\|_* \leq G_p} \langle \boldsymbol{w}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle \right]
$$

$$
P^{(3)} := \mathbb{E}\left[\mathbf{1}_{\mathcal{E}^c} \max_{\boldsymbol{u} \in \mathcal{W}} \sum_{t \in [T] : \|\boldsymbol{g}_t\|_* \leq G_p} \langle \boldsymbol{w}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle \right]
$$

For part $P^{(1)}$, we apply the individual-sequence regret bound $B_{\hat{T}}(G_p)$ of ALG, where $\hat{T}$ is the random number of rounds, for which we have $\mathbb{E}[\hat{T}] \leq pT$. We then drop the indicator and Jensen the expectation inside to get $P^{(1)} \leq B_{pT}(G_p)$. For part $P^{(2)}$, we use that $\mathcal{E}$ and $\mathrm{LCB}_{t-1} < \|\boldsymbol{g}_t\|_*$ imply that $G_{p - 2u_t - 3/(2t) \ln \frac{1}{\delta}} < \|\boldsymbol{g}_t\|_* < G_p$ to find

$$
\mathbf{1}_{\mathcal{E}} \max_{\boldsymbol{u} \in \mathcal{W}} \sum_{t \in [T] : \mathrm{LCB}_{t-1} < \|\boldsymbol{g}_t\|_* \leq G_p} \langle \boldsymbol{w}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle \leq \mathbf{1}_{\mathcal{E}} \sum_{t \in [T] : G_{p - 2u_t - 3/(2t) \ln \frac{1}{\delta}} < \|\boldsymbol{g}_t\|_* \leq G_p} DG_p
$$

Dropping the indicator, taking expectation and using the definition of quantile yields

$$
\begin{aligned}
P^{(2)} &\leq DG_p \sum_{t=1}^{T} \mathbb{P}\left\{ G_{p - 2u_t - 3/(2t) \ln \frac{1}{\delta}} < \|\boldsymbol{g}_t\|_* < G_p \right\} \\
&\leq DG_p \sum_{t=1}^{T} \left(2u_t + 3/(2t) \ln \frac{1}{\delta} \right) \\
&\leq DG_p \left(4\sqrt{2p(1-p)T \ln T} + \frac{13}{3}(\ln T)^2 + 1 \right)
\end{aligned}
$$

Finally, for $P^{(3)}$ we use that the integrand is bounded by TDG_p, and the error probability by $\mathbb{P}(\mathcal{E}^c) \leq \frac{2}{T}$ to find $P^{(3)} \leq 2DG_p$. ∎