Extending Online Robust Mean Estimation: Improved Error Bounds in Less Adversarial Settings

Malav Patel
Department of Computer Science
Georgia Institute of Technology
Atlanta, GA 30332
mpatel636@gatech.edu

Richard Rex
Department of Computer Science
Georgia Institute of Technology
Atlanta, GA 30332
rarockiasamy3@gatech.edu

Faris Qadan
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
fqadan3@gatech.edu

Abstract

Statistical estimation is a fundamental task in applied statistics, having a broad range of application. Of the plethora of quantities which have been studied, none has received more attention than estimation of the first moment of a distribution, i.e. the mean. Research down this avenue has spawned many sub-fields, including those of *robust* and *high-dimensional* statistics. Recently, these fields have attracted interest from computational perspective, with researchers attempting to provide polynomial time algorithms for computing these quantities under adversarial influence. In this work, we aim to understand robust mean estimation in the online setting and extend it to a use-case that provides a tighter bound on the error achieved by the presented algorithm.

1 Introduction

Statistical estimation is a well studied field with broad application to a variety of domains. One of the most fundamental quantities of interest in statistical estimation is an estimate of the *mean*. In the case of a random variable $X \sim p$, the mean is defined as the first moment of the distribution p. The problem is stated simply as,

Q1: Given i.i.d. samples from an unknown distribution p find an estimate of its mean μ^*

The problem definition above relies on the fact that the distribution p belongs to a family of distributions whose first moment is defined and finite. For example, p can belong to the Gaussian family, binomial family, or exponential family. Note that the list is non-exhaustive. As a counterexample, mean estimation is ill defined if p belongs to the family of Cauchy distributions, as this distribution has undefined mean. The literature has extended statistical estimation to the second (i.e. variance Σ) and even higher moments. A natural extension is the following question,

Q2: Given i.i.d. samples of a random variable from an unknown distribution with finite n-th moment, find an estimate of the i-th moment, where $i \leq n$.

However, mean estimation remains the most fundamental and well-studied. In this work we will restrict ourselves to **Q1**. More specifically we are interested in the field of *robust* mean estimation.

Preprint. Under review.

1.1 Robust Statistics

In the field of robust statistics, researchers and practitioners focus on developing methods that remain reliable even in the presence of potentially disruptive or adversarial influences. Simply, robust statistics aims to find methods that insulate estimates from outliers and/or abnormalities. For example, this can arise in an experimental setting where a researcher incorrectly measures a quantity of interest, leading to an inaccurate estimate of a derived quantity. Robust statistics aims to provide a framework that appropriately filters or down-weights the significance of abnormal measurements when estimating a derived quantity. These methods are crucial for applications where data integrity cannot always be assured, yet accurate estimation of statistical quantities remains essential.

1.1.1 Classes of Contamination

There exist 3 main types of contamination as outlined by Blanchet et al. [1]. We enumerate them below.

1. ϵ-contamination : Originally outlined by Huber [6], the adversary contaminates the data generating distribution p with a distribution h such that new data is now sampled from $\hat{p} = (1 - \epsilon)p + \epsilon h$.
2. full neighborhood contamination: Instead of a local neighborhood contamination model as in ϵ contamination, we allow for full neighborhood contamination drawing samples from a distribution $\hat{p}$ that is an ϵ distance from the original distribution p under some distance metric D. Mathematically, $D(\hat{p}, p) \leq \epsilon$.
3. adaptive contamination: Samples are drawn from p but the adversary is allowed to take an ϵ fraction of these samples and corrupt/replace them. Note that this is the strongest form of contamination.

In what follows, adversaries are capable of adaptive contamination.

A simple example of robust statistics is in the case of least squares linear regression where we seek a solution to $\min_{\theta} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|^2$. In ordinary least squares estimator is sensitive to adversarial inputs as all data points are weighted equally in the estimate. However, *weighted* least squares attempts to insert some insulation to protect the estimate from outliers. A weight matrix $\mathbf{W}$ pre-multiplies the data, effectively weighting each data point. The resulting normal equations are $\mathbf{X}^\top \mathbf{W} \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^\top \mathbf{W} \mathbf{y}$. The weighted least squares method is a very early application in the field of robust statistics. However, note that the user must be aware of the identities of corrupted data points to apply a sensible transformation $\mathbf{W}$. In some cases, this information is not available. We now turn to a classic problem where this phenomenon may arise.

A classic problem in this area is **robust mean estimation**. Here, a user seeks to estimate the mean of an i.i.d. sample drawn from an unknown distribution. However, the situation is complicated by the presence of a malicious adversary who has corrupted a fraction of the data points. The user does not know which specific samples are compromised, but they do have knowledge of the approximate proportion of corrupted data.

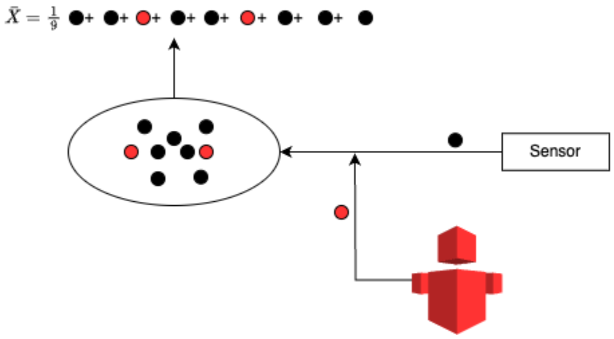

Figure 1: In a simple scenario an adversary corrupts a data stream from which a mean estimate is calculated.

In this challenging setup, the user's objective is to produce an accurate estimate of the true mean, despite the adversarial interference. To achieve this, they must apply techniques that adjust the influence of each sample, essentially weighing the data points based on an assessment of their reliability. By doing so, the user can mitigate the impact of corrupted samples and improve the robustness of the mean estimation, ultimately drawing closer to the true characteristics of the underlying distribution. This line of work has dedicated research towards finding computationally efficient estimators of quantities including but not limited to the mean, covariance, and linear regression in a robust manner [4].

1.2 Computationally Efficient Robust Statistics

While researchers have come up with a variety of methods to solve this problem, a major drawback is that these algorithms often have an exponential run time complexity. This run time is impractical in an application setting where a computation is needed relatively quickly. Take for example, a scenario where a set of distributed sensors provide streams of data from which a mean is calculated. With high dimensional data, an exponential time algorithm will take too long to determine an accurate mean estimate. To address this challenge, a new line of work has looked at the problem from a computational perspective, attempting to design efficient algorithms for statistical estimation [2, 8].

These pioneering works catalyzed a wave of research aimed at developing computationally efficient estimators and enhancing the run time of existing algorithms. A significant body of work has introduced dimension-agnostic algorithms, accompanied by robust theoretical guarantees, for a broad range of complex problems, including linear regression [5] and stochastic optimization [3]. By addressing the formidable challenge posed by high-dimensional data, these algorithms mitigate the curse of dimensionality, offering sub-exponential time solutions for calculating critical quantities in high-dimensional statistics.

1.3 Online Mean Estimation

The works considered thus far operate in the *offline* setting. In other words, an operator collects the entire batch of data before processing it to produce an estimate for the mean. This has its advantages, as all possible data is present to the operator at once and only a single point estimate is produced. Little attention has been paid to the online setting where an operator is tasked with periodically computing an estimate of the mean of a quantity as data is streamed to said operator. Consider the case of n sensors each streaming a scalar at time t to produce a vector of observations. The goal is to provide an estimate of the mean of this vector, taking into consideration the data streamed to the operator from times $< t$. This can be generalized further to an operator receiving n vectors instead of scalars and the resulting objective is to produce a mean estimate of the batch, again taking into account previous batches streamed to the operator. Note that in contrast to the offline case, the operator has very little data from which to create a mean estimate at the beginning of the time horizon. However, as data is streamed the operator is given more information. Only at the end of the time horizon does the operator have access to the data in the same capacity as in the offline setting. Note that in the contamination-free setting, the estimate of the mean at each time does not require knowledge of the sample at other times; an operator can simply take the mean across each dimension to get the mean estimate for the batch.

2 Online Robust Mean Estimation

In this section, we present a detailed exploration of the methodologies and principles applied in the study by Kane et al. [2], focusing particularly on the setup and derivations that culminate in Theorem 6. This theorem is significant as it introduces an efficient, online robust mean estimation algorithm (or more specifically, the key bound that it guarantees). Notably, this algorithm attains an error bound that approaches the optimal offline benchmark, even under adversarial conditions.

2.1 Problem Setup

The basic problem set up in this work contains n sensors, each of which provide d dimensional observations through T rounds. Specifically, $x_t^{(i)}$ captures the batch of coordinates revealed at the t-th round by the the i-th sensor. See Figure 2 for an illustration of this process. In each round t the goal is to output a mean estimate $\mu_t \in \mathbb{R}^d$ of the batch. Note that in the adversarial setting, information from previous batches becomes useful when estimating the current batch mean.

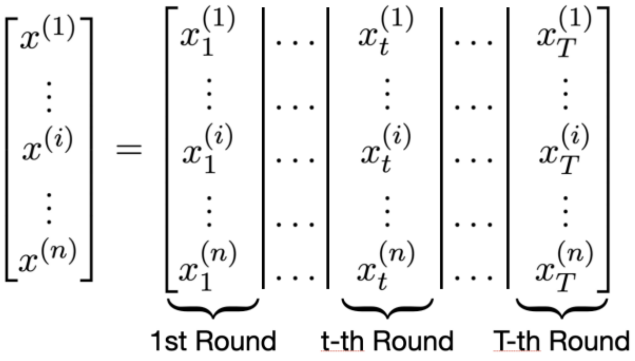

Figure 2: Problem setup. Image from Kane et al. [7]

In accordance with the authors' characterization, online robust mean estimation is structured here under a standard strong contamination model. Formally, consider a set $\mathcal{C}$ of n samples and some parameter $0 \le \epsilon < 1/2$. After observing the entire set $\mathcal{C}$, a *strong contamination adversary* can remove up to ϵn samples from $\mathcal{C}$ and replace them with arbitrary points. This results in an ϵ*-corrupted* version of $\mathcal{C}$, given by $\mathcal{X}$. We note this form of contamination is as defined above (see Section 1.1.1).

In the context of online robust mean estimation specifically, we are given the same set $\mathcal{X} = \{x^{(1)}, \dots, x^{(n)}\}$. Here, the samples are distributed on some $\mathbb{R}^M$ with unknown mean vector $\mu^* \in \mathbb{R}^M$. The aim then becomes to estimate μ^* 'online', over T rounds of observed data, without the added benefit of seeing the entire set in one go as is customary in the offline setting.

2.1.1 Formal Guarantees

At each round $t \in [T]$, the algorithm observes a new batch of coordinates from each samples, given by $x^{(i)} \in \mathbb{R}$ for $i \in [n]$, with some parameters $d \cdot T = M$. By round T, the algorithm is said to estimate the mean of X under ϵ-corruption in the T-round online setting with error $\epsilon' > 0$, failure probability $\tau \in (0, 1)$ and sample complexity n, if with probability (w.p.) at least $1 - \tau$,

$$\|\mu - \mu^*\|_2 = \sqrt{\sum_{t=1}^{T} \|\mu_t - \mu_t^*\|_2^2} \le \epsilon'$$

2.1.2 Stability Conditions

The clean sample $\mathcal{C}$ as defined by the authors also satisfies the notion of (ϵ, δ)-stability. Formally, for $\epsilon \in (0, \frac{1}{2}$ and $\delta \ge \epsilon$, $\mathcal{C} \subset \mathbb{R}^d$ is (ϵ, δ)-stable with respect to $\mu \in \mathbb{R}^d$ if and only if

1. Deviations of the mean along any v are bounded: $\left|\frac{1}{|C'|} \sum_{x \in C'} v^T(x - \mu)\right| \le \delta$
2. Deviations of the variance along any v are controlled: $\left|\frac{1}{|C'|} \sum_{x \in C'} (v^T(x - \mu))^2) - 1\right| \le \frac{\delta^2}{\epsilon}$

for all $v \in \mathbb{R}^d$ and every subset $C' \subset C$ where $\left|C'\right| \ge (1 - \epsilon)|C|$. This plays a critical role in limiting the effect of adversarial samples and producing a reliable estimate for μ within controlled bounds.

2.2 *Efficient* Online Robust Mean Estimation

Theorem 1 [2] provides an efficient algorithm for online robust mean estimation under the (ϵ, δ)-stability assumption. Specifically, the algorithm promises two favorable guarantees:

- An l_2 error bound of $O(\delta \log T)$
- A polynomial runtime (poly-time) with respect to sample size (n) and dimension (M)

Concretely, these guarantees are underpinned by (a) its stability assumption, and (b) its use of a *weighted filtering approach*. These will prove crucial in our analysis and furthering of the authors' work.

2.2.1 (ϵ, δ)-Stability for Recursive Error Control

The stability assumption is fundamental as it ensures that clean samples dominate the estimation process, even in the presence of adversarial outliers. Algorithm 1 [2] leverages this property to effectively give adversarial samples lower weight, while retaining sufficient information from the clean samples for accurate estimation.

Formally, the stability condition guarantees that the deviation of the empirical covariance matrix from the identity is bounded, specifically $\|\Sigma - I\|_2 \leq O(\frac{\delta^2}{\epsilon})$. At each round t, the algorithm updates the weighted covariance matrix Σ to satisfy $\|\Sigma\|_2 \leq 1 + \lambda$, where $\lambda = O(\frac{\delta^2}{\epsilon})$. This ensures that the influence of clean samples remains dominant, enabling the mean estimation to converge towards the true mean within a controlled number of rounds.

2.2.2 Weighted Online Filtering

The algorithm below [7] outlines an online filter that iteratively updates its "trust" in each of its sources, as presented by the authors.

Algorithm 1 Online Filter

```
Require: Number of samples $n$, Byzantine fraction $\epsilon$, round number $T$, sample coordinates $x_t^{(i)}$ revealed at the $t$-th round for $t = 1, 2, \ldots, T$, filter threshold $\lambda$, and initialized weight $w_0^{(i)} = 1$, for $i = 1, 2, \ldots, n$.
for $t = 1, 2, \ldots, T$ do
    Initialize $w_t \leftarrow w_{t-1}$ and update $\bar{x}_t^{(i)} = \left(x_t^{(i)}, \ldots, x_t^{(i)}\right)$.
    Compute $\Sigma \leftarrow$ WCov$(w_t, \bar{x}_t^{(1:n)})$.
    while $\|\Sigma\|_2 > 1 + \lambda$ do
        Compute the top eigenvector $v$ of $\Sigma$.
        Compute empirical weighted mean $\mu(w_t) \leftarrow \sum_{i=1}^n w_t^{(i)} \bar{x}_t^{(i)}$.
        for $i = 1, 2, \ldots, n$ do
            Compute $\rho^{(i)} \leftarrow \langle v, x^{(i)} - \mu(w_t) \rangle^2$.
        end for
        Sort $\rho^{(1:n)}$ into a decreasing order denoted as $\rho^{(1)} \geq \rho^{(2)} \geq \cdots \geq \rho^{(n)}$, and let $\beta$ be the smallest number such that $\sum_{i=1}^{\beta} \rho^{(i)} > 2\epsilon$.
        Apply WFilter to update weights for $\{\pi(1), \ldots, \pi(\beta)\}$ as $\{w_t^{(1)}, \ldots, w_t^{(\beta)}\} \leftarrow$ WFilter$(\rho^{(1:\beta)}, w_t^{(1:\beta)})$; while the remaining weights stay the same, i.e., $w_t^{(i)} = w_t^{(i)}$ for $i > \beta$.
        Update $\Sigma \leftarrow$ WCov$(w_t, \bar{x}_t^{(1:n)})$.
    end while
    $\mu_t \leftarrow \sum_{i=1}^n w_t^{(i)} x_t^{(i)}$.
end for
Ensure: $\mu_t$.
```

Weighted Filter (WFilter)

```
Require: Scores $\rho^{(1:\beta)}$, weights $w^{(1:\beta)}$.
  for $i = 1, 2, \ldots, \beta$ do
      $w_t^{(i)} \leftarrow \left(1 - \frac{\rho^{(i)}}{\max_j \rho^{(j)}}\right) w_t^{(i)}$.
  end for
Ensure: $w^{(1:\beta)}$.
```

```
Weighted Covariance (WCov)
Require: Weight $w = w^{(1:n)}$ and samples $\bar{x}^{(1:n)}$.
  Compute $\mu(w) \leftarrow \sum_{i=1}^{n} \frac{w^{(i)}}{\|w\|_1} \bar{x}^{(i)}$.
  Compute weighted covariance estimation $\Sigma \leftarrow \sum_{i=1}^{n} \frac{w^{(i)}}{\|w\|_1} \left(\bar{x}^{(i)} - \mu(w)\right) \left(\bar{x}^{(i)} - \mu(w)\right)^{\top}$.
Ensure: $\Sigma$.
```

Algorithm 1 is an online filtering procedure designed to iteratively refine its characterization of each sample received (or rather, its confidence in the reliability of each sample). This is achieved by maintaining and dynamically updating a set of weights representing a ‘current’ assessment of whether or not a sample is corrupted.

In short, the algorithm computes a weighted mean and covariance between each of the observed sample blocks at each round t, reducing the weight of samples disproportionately contributing to the observed instability.

This algorithm guarantees Theorem 1’s error bound by two key observations:

1. Since the covariance matrix after each filtering step satisfies $\left\|\Sigma\right\|_2 \leq 1 + \lambda$, the error in the weighted mean can be bounded by $\left\|\mu(w_t) - \mu^*\right\|_2 \leq O(\delta + \sqrt{\epsilon\lambda})$.
2. The difference in estimates of μ_t^* with weights from two consecutive rounds is proportional to the difference in their weights; this implies that $\left\|\mu(w_t) - \mu(w_{t+1})\right\|_2^2 \leq O(\frac{\delta^2}{\epsilon}) \cdot \left\|w_t - w_{t+1}\right\|_1$.

Combining these recursively, we finally get that $\sum_{t=1}^{T} \|\mu_t - \mu^*\|_2^2 \leq O(\delta^2 \text{log}^2 T)$; a bound on the estimation error given our assumed stability. These proofs are outlined by the authors in Lemmas 3 and 4, leading to Theorem 6; they are simply outlined here for the sake of brevity, and are examined in greater depth in Section 3 below.

As such, this algorithm, in conjunction with the stability setup, is key in ensuring that the cumulative error remains bounded by $O(\delta \text{log} T)$ while maintaining a polynomial runtime.

3 Extension

One of the main findings of Kane et al. [7] involves a bound on the error of the online mean estimate, as computed by the filtering-based algorithm described in Section 2.

This result provides a robust guarantee for online mean estimation but hinges on assumptions that lead to a ‘conservative’ bound. Specifically, the growth of the error bound arises from the unconstrained nature of the weight differences $\left\|w_t - w_{t'}\right\|_1$ used in the recursive decomposition of error. While the filtering algorithm reduces adversarial influence over time, the lack of explicit control over the evolution of weights introduces pessimism in the cumulative error analysis.

In this section, we motivate the need for an improved error bound, introduce a formal bound on the decay weights across rounds, and use this to prove a new bound on the algorithm’s error.

3.1 Motivation

Kane et al. [7] provide a robust analysis of online mean estimation under contaminated, adversarial settings. However, the authors make a critical assumption that leads to conservative error bounds. Specifically, the recursive decomposition of error depends on the weight difference $\|w_t - w_{t'}\|_1$ between two rounds t and t' (where $1 \leq t < t' \leq T$). The authors further show proportionality between weight differences and the mean difference (or more formally, $\left\|\mu(w_t, x_{1:t}) - \mu(w_{t'}, x_{1:t})\right\|_2^2 \propto \|w_t - w_{t'}\|_1$).

However, we observe that they do not impose any explicit constraint on how $\|w_t - w_{t'}\|_1$ evolves over time. This lack of a decay model allows $\|w_t - w_{t'}\|_1$ to grow logarithmically as rounds progress. When summed recursively over all pairs of rounds, this leads to a pessimistic squared error bound of the order of $O(\log^2 T)$.

Consider an environment in which adversarial impact diminishes over time, or where data is relatively stable for long periods of time with minimal adversarial intervention. One such example involves federated learning (FL) for predictive text models on phones. In these cases, learning is locally (differentially) private, and stabilizes to a highly predictable state as users interact more with their personal systems. Notably, the potential for adversarial harm here is quite low. Other distributed prediction models boast similar properties, where adversarial impact diminishes drastically over time (or at least, is highly controlled).

In cases such as this, one may look to opt for a stricter error bound for more predictable deviations from the true mean estimate. This is especially true as with minimal adversarial impact, many of these cases are likely to enjoy estimates that come close to the true mean.

3.2 Bounding Weight Decay Over Time

Considering the motivation above, the question remains of formally defining the bound on the weight decay. Note first that the bound must be strictly decreasing with the round t. Furthermore, to encapsulate the property of a “diminishing" adversary, we require that the change in the upper bound with respect to t also diminishes. An appropriate choice is a reciprocal time decay of the form $\frac{1}{t}$. More specifically, we bound the $\ell 2$-norm of the weight change between two time steps t and t' using this reciprocal time dependence:

$$\|w_t - w_{t'}\|_1 \leq O(\frac{1}{t - t' + 1})$$

The inequality above effects a diminishing influence by the adversary.

3.3 Reformulating *Lemma 3* [7]

Under the assumption that the weight decay is bounded as in section 3.2, we get a tighter bound on the mean shift between any two rounds t, t'.

$$\left\|\mu\left(w_t, \bar{x}_{(1:n)}\right) - \mu\left(w_{t'}, \bar{x}_{(1:n)}\right)\right\|_2^2 \leq O(1) \cdot \frac{\delta^2}{\epsilon} \cdot O(\frac{1}{t - t' + 1})$$

To establish the proof, we follow a similar vein to the proof of Lemma 3 in Kane at al. [7]. Consider the normalized categorical distributions y_t and y_t' derived from w_t and w_t' at two different rounds t and t' respectively. Define $\eta = \|y_t - y_{t'}\|_1$. Now consider writing y_t as a η-mixture of y_t' and another categorical distribution e (i.e. $\|e\|_1 = 1$). Specifically, $y_t = (1-\eta)y_t' + \eta e$. Define $\mathrm{Cov}(\cdot)$ and $\mu(\cdot)$ as the covariance and mean over the batch mean $\bar{x}_t^{(i)}$, which is an estimator of the true batch mean. Then we have,

$$\mathrm{Cov}(y_t) = (1-\eta) \cdot \mathrm{Cov}(y_{t'}) + \eta \cdot \mathrm{Cov}(e) + \eta(1-\eta) \cdot (\mu(y_{t'}) - \mu(e))(\mu(y_{t'}) - \mu(e))^\top$$

Since both $\|\mathrm{Cov}(y_t)\|_2 \leq 1 + O(\frac{\delta^2}{\epsilon})$ and $\|\mathrm{Cov}(y_{t'})\|_2 \leq 1 + O(\frac{\delta^2}{\epsilon})$ by Lemma 1 (see Kane et al. [7]), we have that $\|\mu(y_{t'}) - \mu(e)\|_2^2 = O(\delta^2/(\eta\epsilon))$. Note that we can rewrite the mean difference between y_t' and e as a mean difference between y_t and y_t' which will be useful: $\mu(y_t) - \mu(y_{t'}) = (1-\eta) \cdot \mu(y_{t'}) + \eta \cdot \mu(e) - \mu(y_{t'}) = -\eta \cdot (\mu(y_{t'}) - \mu(e))$. Namely, the norm of the differences is scaled by a factor $\eta = \|y_t - y_{t'}\|_1 = O(\|w_t - w_t'\|_1) \leq O(\frac{1}{t-t'+1})$.

Using this fact we have $\|\mu(y_t) - \mu(y_{t'})\|_2^2 \leq O(\eta \cdot \delta^2/\epsilon) \leq O(\frac{\delta^2}{\epsilon}) \cdot O(\frac{1}{t-t'+1})$. Note the factor of η arises in the numerator because the norms are *squared*. Finally, notice that by the same definition as the original Lemma, $\mu(y_t)$ is equivalent to $\mu(w_t, \bar{x}_t^{(1:n)})$. Rewriting, we find

$$\begin{aligned}
\|y_t - y_{t'}\|_2^2 &= \left\|\mu\left(w_t, \bar{x}_{(1:n)}\right) - \mu\left(w_{t'}, \bar{x}_{(1:n)}\right)\right\|_2^2 \\
&\leq O(\frac{\delta^2}{\epsilon}) \cdot O(\frac{1}{t - t' + 1}) \\
&= O(1) \cdot \frac{\delta^2}{\epsilon} \cdot O(\frac{1}{t - t' + 1})
\end{aligned}$$

where in the second to third lines, we separate the constant factor multiplying $\frac{\delta^2}{\epsilon}$ into an $O(1)$ factor. Thus we have shown that the adversary has diminishing impact on mean estimates for rounds further in the future.

3.4 Reformulating *Lemma 4* [7]

With the refined bound established above in the 're-proof' of Lemma 3, we can now derive a new bound on the error proposed in Lemma 4.

Under the assumption of weight decay introduced in 3.2, we look to update the cumulative l_2-error of the online robust mean estimation algorithm, given by $\sum_{t=1}^{T} \left\| \mu(w_t, \bar{x}^{(1:n)}) - \mu(w_{t'} - \bar{x}^{(1:n)}) \right\|_2^2$. By triangle inequality, we first bound this value by the following:

$$\sum_{t=1}^{T} \left\| \mu(w_t, \bar{x}^{(1:n)}) - \mu(w_{t'} - \bar{x}^{(1:n)}) \right\|_2^2 \leq \sum_{t=1}^{T} \sum_{t'<t} \left\| \mu(w_t, \bar{x}^{(1:n)} - \mu(w_{t'}, \bar{x}^{(1:n)}) \right\|_2^2$$

As such, we are able to substitute the result from 3.3, to further bound this:

$$\sum_{t=1}^{T} \left\| \mu(w_t, \bar{x}^{(1:n)}) - \mu(w_{t'} - \bar{x}^{(1:n)}) \right\|_2^2 \leq \sum_{t=1}^{T} \sum_{t'<t} O(1) \cdot \frac{\delta^2}{\epsilon} \cdot \frac{1}{t - t' + 1}$$

$$= O(1) \cdot \frac{\delta^2}{\epsilon} \sum_{t=1}^{T} \sum_{t'<t} \frac{1}{t - t' + 1}$$

We consider the inner summation over t', which can be simply evaluated with a change of variable; $k := t - t' + 1$, such that

$$\sum_{t'<t} \frac{1}{t - t' + 1} = \sum \frac{1}{k}$$

We note that this represents a harmonic series, and thus, $\sum_{k=1}^{t} \frac{1}{k} = O(\text{log}t)$ for some fixed t. Note that for any length t interval starting after 1 has its harmonic sum bounded above by $O(\log t)$. Using this fact we can plug this back into our originally obtained bound:

$$\sum_{t=1}^{T} \left\| \mu(w_t, \bar{x}^{(1:n)}) - \mu(w_{t'} - \bar{x}^{(1:n)}) \right\|_2^2 \leq O(1) \cdot \frac{\delta^2}{\epsilon} \sum_{t=1}^{T} O(\text{log}t)$$

Again, considering the sum $\sum_{t=1}^{T} O(\text{log}t)$, we arrive simply at $O(T\text{log}T - T) = O(T\text{log}T)$. This is simply an evaluation of an integral of the natural logarithm of t, in the discrete case.

Finally, we are able to substitute this into our original equation to arrive at the following:

$$\sum_{t=1}^{T} \left\| \mu(w_t, \bar{x}^{(1:n)}) - \mu(w_{t'} - \bar{x}^{(1:n)}) \right\|_2^2 \leq O(1) \cdot \frac{\delta^2}{\epsilon} \cdot O(T\text{log}T)$$

$$= O(\frac{T\delta^2}{\epsilon} \cdot \text{log}T)$$

Notably, this result is a stark improvement from Lemma 4 as presented by the authors, who present a square log-bound on this quantity, as opposed to the log-bound derived above.

3.5 Reformulating *Theorem 6* [7]

To restate it concisely, Theorem 6 (in its original form) dictates that using Algorithm 1, the sequence of mean estimates follows a bound of the form

$$\|\mu - \mu^*\|_2 \leq O(\delta \text{log}T)$$

However, this bound hinges heavily on the results of Lemmas 3 and 4. As such, we can reformulate it given the updated bounds developed above.

To begin, we note that the error bound can be split into a recursive component and an offline one. The latter, as stated by the authors, remains bounded $\left\|\mu(w_T, \bar{x}_T) - \mu^*\right\|_2^2 \leq O(\delta^2)$.

The recursive error bound differs slightly, as given by Lemma 4 above:

$$\sum_{t=1}^{T} \left\|\mu(w_t, \bar{x}^{(1:n)}) - \mu(w_{t'} - \bar{x}^{(1:n)})\right\|_2^2 \leq O(\frac{T\delta^2}{\epsilon} \cdot \log T)$$

As noted previously, Theorem 6 combines these two quantities by the triangle inequality, such that

$$\|\mu - \mu^*\|_2^2 \leq O(\delta^2) + O(\frac{T\delta^2}{\epsilon} \cdot \log T)$$

$$= O(\frac{T\delta^2}{\epsilon} \cdot \log T)$$

Above, the recursive error dominates over an increasing number of rounds. As such, we take the square root of both sides to arrive at a finalized update on Theorem 6 given by:

$$\|\mu - \mu^*\|_2 \leq O(\delta \cdot \sqrt{\frac{T}{\epsilon}} \cdot \sqrt{\log T})$$

More concisely, this improves on the error bound proposed by the authors with some additional constants, and a log-bound on the squared error bound, as opposed a to a squared log one.

4 Conclusion

In this work, we extended the framework of online robust mean estimation as proposed by Kane et al. [7]. Online robust mean estimation requires navigating a nuanced trade-off between optimism and realism to arrive at an estimate; the authors of this work chose to adopt a relatively conservative estimation strategy. Motivated by real-world applications of online learning, we introduced a novel – but general – constraint on the rate of weight decay over time, recognizing its critical influence on the proportional relationship between mean decay and weight decay observed in the authors' framework.

Building upon this insight, we revisited and reformulated two key lemmas from the original work, incorporating our weight decay constraint to broaden the horizon of this work to a greater line of applications. This enabled us to propose a tighter bound on the error of the estimated mean across multiple rounds of learning. Specifically, under conditions of diminishing adversarial influence, we demonstrated that an error bound of $O(\sqrt{T \log T})$ can be achieved, improving upon prior results and offering stronger guarantees for robust learning.

Our results hold significant implications in an era where computational efficiency and rapid training are paramount. As such, the introduction of a general framework for incorporating weight decay constraints into online robust mean estimation opens avenues for further research. By formalizing bounds on weight decay tailored to specific application contexts, this framework has the potential to enhance robustness and accuracy in diverse settings. These contributions not only advance the theory of online robust mean estimation but also lay the groundwork for practical implementations in resource-constrained environments.

5 Future Work

As was touched on before, the immediate extension to our work would be to look into more application-specific bounds on weight decay. This can be done empirically or mathematically, depending on the application. The bound proposed above serves as a general framework for expanding on the work of Kane et al. [7] for less adversarial systems, rather than a finalized bound to be used in all applications.

Furthermore, the same sort of expansions can be performed for the online estimation of higher-order moments in these less adversarial settings, as initially proposed by the authors in the paper.

References

[1] Jose Blanchet, Jiajin Li, Sirui Lin, and Xuhui Zhang. Distributionally robust optimization and robust statistics, 2024.

[2] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[3] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019.

[4] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.

[5] 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, 2019.

[6] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[7] Daniel M. Kane, Ilias Diakonikolas, Hanshen Xiao, and Sihan Liu. Online robust mean estimation, 2023.

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