# Robust Stochastic Optimization via Gradient Quantile Clipping

**Anonymous authors**
**Paper under double-blind review**

## Abstract

We introduce a clipping strategy for Stochastic Gradient Descent (SGD) which uses quantiles of the gradient norm as clipping thresholds. We prove that this new strategy provides a robust and efficient optimization algorithm for smooth objectives (convex or non-convex), that tolerates heavy-tailed samples (including infinite variance) and a fraction of outliers in the data stream akin to Huber contamination. Our mathematical analysis leverages the connection between constant step size SGD and Markov chains and handles the bias introduced by clipping in an original way. For strongly convex objectives, we prove that the iteration converges to a concentrated distribution and derive high probability bounds on the final estimation error. In the non-convex case, we prove that the limit distribution is localized on a neighborhood with low gradient. We propose an implementation of this algorithm using rolling quantiles which leads to a highly efficient optimization procedure with strong robustness properties, as confirmed by our numerical experiments.

## Contents

## 1 Introduction

Stochastic gradient descent (SGD) (Robbins & Monro, 1951) is the core optimization algorithm at the origin of most stochastic optimization procedures (Kingma & Ba, 2014; Defazio et al., 2014; Johnson & Zhang, 2013). SGD and its variants are ubiquitously employed in machine learning in order to train most models (Kushner & Yin, 2003; Benveniste et al., 2012; Lan, 2020; Shalev-Shwartz et al., 2007; Bottou et al., 2018; Ma et al., 2018). The convergence properties of SGD are therefore subjects of major interest.

Early studies of SGD convergence generally relied on strong assumptions such as bounded domain (Shalev-Shwartz et al., 2009) or uniformly bounded gradient variance (Rakhlin et al., 2011) and obtained error bounds in expectation. With the recent resurgence of interest for robust statistics (Hsu & Sabato, 2016; Diakonikolas et al., 2019; Lecué & Lerasle, 2017; Prasad et al., 2018), variants of SGD based on clipping are shown to be robust to heavy-tailed gradients (Gorbunov et al., 2020; Tsai et al., 2022), where the gradient samples are only required to have a finite variance. The latter requirement has been further weakened to the existence of a $q$-th moment for some $q > 1$ in (Sadiev et al., 2023; Nguyen et al., 2023). In this paper, we go further and show that another variant of clipped SGD with proper thresholds is robust both to heavy tails *and* outliers in the data stream.

Robust statistics appeared in the 60s with the pioneering works of Huber, Tukey and others (Tukey, 1960; Huber, 1992; 1972; Rousseeuw & Hubert, 2011; Hampel, 1971). More recently, the field found new momentum thanks to a series of works about robust scalar mean estimation (Catoni, 2012; Alon et al., 1996; Jerrum et al., 1986; Lugosi & Mendelson, 2021) and the more challenging multidimensional case (Hopkins, 2020; Catoni & Giulini, 2018; Lugosi & Mendelson, 2019; Minsker, 2015; Cherapanamjeri et al., 2019; Depersin & Lecué, 2022; Lei et al., 2020; Diakonikolas et al., 2020). These paved the way to the elaboration of a host of robust learning algorithms (Holland & Ikeda, 2019; Prasad et al., 2018; Lecué & Lerasle, 2017; Liu et al., 2020; Pensia et al., 2020) which have to date overwhelmingly focused on the batch learning setting. We consider the setting of streaming stochastic optimization (Bottou & Cun, 2003; Bottou & Lecun, 2005; McMahan et al., 2013), which raises an additional difficulty coming from the fact that algorithms can see each sample only once and must operate under an $\mathcal{O}(d)$ memory and complexity constraint for $d$-dimensional optimization problems. A limited number of papers (Tsai et al., 2022; Nazin et al., 2019; Diakonikolas et al., 2022) propose theoretical guarantees for robust algorithms learning from streaming data.

This work introduces such an algorithm that learns from data on the fly and is robust both to heavy tails and outliers, with minimal computational overhead and sound theoretical guarantees.

We consider the problem of minimizing a smooth objective

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta) := \mathbb{E}_\zeta[\ell(\theta,\zeta)] \tag{1}$$

using observations $G(\theta,\zeta_t)$ of the unknown gradient $\nabla\mathcal{L}(\theta)$, based on samples $(\zeta_t)_{t\geq 0}$ received sequentially that include corruptions with probability $\eta < 1/2$. Formulation (1) is common to numerous machine learning problems where $\ell$ is a loss function evaluating the fit of a model with parameters $\theta$ on a sample $\zeta$, the expectation $\mathbb{E}$ is w.r.t the unknown uncorrupted sample distribution.

We introduce quantile-clipped SGD (QC-SGD) which uses the iteration

$$\theta_{t+1} = \theta_t - \alpha_{\theta_t}\beta G(\theta_t,\zeta_t) \quad \text{with} \quad \alpha_{\theta_t} = \min\Big(1, \frac{\tau_{\theta_t}}{\|G(\theta_t,\zeta_t)\|}\Big), \tag{2}$$

where $\beta > 0$ is a constant step size and $\alpha_{\theta_t}$ is the clipping factor with threshold chosen as the $p$-th quantile $\tau_{\theta_t} = Q_p(\|\widetilde{G}(\theta_t,\zeta_t)\|)$ with $\widetilde{G}(\theta_t,\zeta_t)$ an uncorrupted sample of $\nabla\mathcal{L}(\theta_t)$ and $p\in(0,1)$ (details will follow). Quantiles are a natural choice of clipping threshold which allows to handle heavy tails (Rothenberg et al., 1964; Bloch, 1966) and corrupted data. For instance, the trimmed mean offers a robust and computationally efficient estimator of a scalar expectation (Lugosi & Mendelson, 2021). Since the quantile $Q_p(\|\widetilde{G}(\theta_t,\zeta_t)\|)$ is non-observable, we introduce a method based on rolling quantiles in Section 5 which keeps the procedure $\mathcal{O}(d)$ both memory and complexity-wise. The main benefit of QC-SGD 2 is to grant robustness to the presence of a proportion $\eta < 1/2$ of corruptions in the stream of gradient samples. This could not be achieved by previous clipped SGD methods (Gorbunov et al., 2020; Tsai et al., 2022; Sadiev et al., 2023; Nguyen et al., 2023). We also show that iteration (2) is adaptive to heavy-tailed gradient variance and converges to a limit distribution with strong concentration properties.

**Contributions.** Our main contributions are as follows:

- For small enough $\eta$ and well-chosen $p$, we show that, whenever the optimization objective is smooth and strongly convex, QC-SGD converges *geometrically* to a limit distribution such that the deviation around the optimum achieves the *optimal* dependence on $\eta$.
- In the non-corrupted case $\eta = 0$ and with a strongly convex objective, we prove that a coordinated choice of $\beta$ and $p$ ensures that the limit distribution is sub-Gaussian with constant of order $\mathcal{O}(\sqrt{\beta})$. In the corrupted case $\eta > 0$, the limit distribution is sub-exponential.
- For a smooth objective (non-convex) whose gradient satisfies an identifiability condition, we prove that the total variation distance between QC-SGD iterates and its limit distribution vanishes sub-linearly. In this case, the limit distribution is such that the deviation of the objective gradient is optimally controlled in terms of $\eta$.
- Finally, we provide experiments to demonstrate that QC-SGD can be easily and efficiently implemented by estimating $Q_p(\|\widetilde{G}(\theta_t,\zeta_t)\|)$ with rolling quantiles. In particular, we show that the iteration is indeed robust to heavy tails and corruption on multiple stochastic optimization tasks.

Our theoretical results are derived thanks to a modelling through Markov chains and hold under an $L_q$ assumption on the gradient distribution with $q > 1$.

**Related works.** Convergence in distribution of the Markov chain generated by constant step size SGD, relatively to the Wasserstein metric, was first established in (Dieuleveut et al., 2020). Another geometric convergence result was derived in (Yu et al., 2021) for non-convex, non-smooth, but quadratically growing objectives, where a convergence statement relatively to a weighted total variation distance is given and a CLT is established. These papers do not consider robustness to heavy tails or outliers. Early works proposed stochastic optimization and parameter estimation algorithms which are robust to a wide class of noise of distributions (Martin & Masreliez, 1975; Polyak & Tsypkin, 1979; 1981; Price & VandeLinde, 1979; Stanković & Kovačević, 1986; Chen et al., 1987; Chen & Gao, 1989; Nazin et al., 1992), where asymptotic convergence guarantees are stated for large sample sizes. Initial evidence of the robustness of clipped SGD to heavy tails was given by (Zhang et al., 2020) who obtained results in expectation. Subsequent works derived high-confidence sub-Gaussian performance bounds under a finite variance assumption (Gorbunov et al., 2020; Tsai et al., 2022) and later under an $L_q$ assumption (Sadiev et al., 2023; Nguyen et al., 2023) with $q > 1$. A similar SGD clipping scheme to (2) is presented in (Seetharaman et al., 2020), however, in contrast to our work, they do not consider the robust setting and focus on experimental study while we also provide theoretical guarantees.

Robust versions of Stochastic Mirror Descent (SMD) are introduced in (Nazin et al., 2019; Juditsky et al., 2023). For a proper choice of the mirror map, SMD is shown to handle infinite variance gradients without any explicit clipping (Nemirovskij & Yudin, 1983; Vural et al., 2022). Finally, (Diakonikolas et al., 2022) study heavy-tailed and outlier robust streaming estimation algorithms of the expectation and covariance. On this basis, robust algorithms for linear and logistic regression are derived. However, the involved filtering procedure is hard to implement in practice and no numerical evaluation of the considered approach is proposed.

**Agenda.** In Section 2 we set notations, state the assumptions required by our theoretical results and provide some necessary background on continuous state Markov chains. In Section 3, we state our results for strongly convex objectives including geometric ergodicity of QC-SGD (Theorem 1), characterizations of the limit distribution and deviation bounds on the final estimate. In Section 4, we remove the convexity assumption and obtain a weaker ergodicity result (Theorem 2) and characterize the limit distribution in terms of the deviations of the objective gradient. Finally, we present a rolling quantile procedure in Section 5 and demonstrate its performance through a few numerical experiments on synthetic and real data.

## 2 Preliminaries

The model parameter space is $\mathbb{R}^d$ endowed with the Euclidean norm $\|\cdot\|$, $\mathcal{B}(\mathbb{R}^d)$ is the Borel $\sigma$-algebra of $\mathbb{R}^d$ and we denote by $\mathcal{M}_1(\mathbb{R}^d)$ the set of probability measures over $\mathbb{R}^d$. We assume throughout the paper that the objective $\mathcal{L}$ is smooth.

**Assumption 1.** *The objective $\mathcal{L}$ is $L$-Lipschitz-smooth, namely*

$$\mathcal{L}(\theta') \le \mathcal{L}(\theta) + \langle \nabla\mathcal{L}(\theta), \theta' - \theta\rangle + \frac{L}{2}\|\theta - \theta'\|^2$$

*with $L < +\infty$ for all $\theta, \theta' \in \mathbb{R}^d$.*

The results from Section 3 below use the following

**Assumption 2.** *The objective $\mathcal{L}$ is $\mu$-strongly convex, namely*

$$\mathcal{L}(\theta') \ge \mathcal{L}(\theta) + \langle \nabla\mathcal{L}(\theta), \theta' - \theta\rangle + \frac{\mu}{2}\|\theta - \theta'\|^2$$

*with $\mu > 0$ for all $\theta, \theta' \in \mathbb{R}^d$.*

An immediate consequence of Assumption 2 is the existence of a unique minimizer $\theta^\star = \arg\min_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta)$. The next assumption formalizes our corruption model.

**Assumption 3** ($\eta$-corruption)**.** *The gradients $(G(\theta_t, \zeta_t))_{t\ge 0}$ used in Iteration (2) are sampled as $G(\theta_t, \zeta_t) = U_t\breve{G}(\theta_t) + (1 - U_t)\widetilde{G}(\theta_t, \zeta_t)$ where $U_t$ are i.i.d Bernoulli random variables with parameter $\eta < 1/2$, $\breve{G}(\theta_t) \sim \mathcal{D}_{\mathcal{O}}(\theta_t)$ with $\mathcal{D}_{\mathcal{O}}(\theta_t)$ an arbitrary distribution and $\widetilde{G}(\theta_t, \zeta_t) \sim \mathcal{D}_{\mathcal{I}}(\theta_t)$ follows the true gradient distribution and is independent from the past given $\theta_t$.*

Assumption 3 is an online analog of the Huber contamination model (Huber, 1965; 1992) where corruptions occur with probability $\eta$ and where the distribution of corrupted samples $\mathcal{D}_{\mathcal{O}}(\theta_t)$ (outliers) is not fixed and may depend on the current iterate $\theta_t$. On the other hand, $\mathcal{D}_{\mathcal{I}}(\theta_t)$ denotes the distribution of inliers. This notation and dichotomy between inliers and outliers follows the example of (Lecué & Lerasle, 2017; Lecué &

Lerasle, 2019). Assumption 3 corresponds to *additive* contamination (Diakonikolas & Kane, 2023, Section 1.2.2) where corruptions are only added to the data. A more general TV-contamination model allowing for true samples to be adversely removed is used in (Diakonikolas et al., 2022). Note however, that the latter mainly focuses on mean estimation. Note also that additive contamination remains realistic since it accounts for invalid entries occurring in the data stream even if it doesn't support entries being targeted and censored. The next assumption requires the true gradient distribution to be unbiased and diffuse.

**Assumption 4.** *For all $\theta$, non-corrupted gradient samples $\widetilde{G}(\theta,\zeta) \sim \mathcal{D}_{\mathcal{I}}(\theta)$ are such that*

$$\widetilde{G}(\theta,\zeta) = \nabla\mathcal{L}(\theta) + \varepsilon_\theta, \tag{3}$$

*where $\varepsilon_\theta$ is a centered noise $\mathbb{E}[\varepsilon_\theta|\theta] = 0$ with distribution $\delta\nu_{\theta,1} + (1-\delta)\nu_{\theta,2}$ where $\delta > 0$ and $\nu_{\theta,1}, \nu_{\theta,2}$ are distributions over $\mathbb{R}^d$ such that $\nu_{\theta,1}$ admits a density $h_\theta$ w.r.t. the Lebesgue measure satisfying*

$$\inf_{\|\omega\|\le R} h_\theta(\omega) > \varkappa(R) > 0$$

*for all $R > 0$, where $\varkappa(\cdot)$ is independent of $\theta$.*

In addition to the unbiased property, Assumption 4 imposes that the noise distribution be expressible as the combination of two components, one of which must be diffuse with density satisfying a minorization inequality. Note that this is a weak constraint since it is satisfied, for example, as soon as the noise $\varepsilon_\theta$ admits a density w.r.t. Lebesgue's measure which is positive everywhere. This condition is similar to (Yu et al., 2021, Assumption 2.3) since both find their origin in Markov chain minorization conditions (Meyn & Tweedie, 1993, Section 5.2). These ensure that a chain properly explores its state space and are a common way to prove Markov chain convergence (Rosenthal, 1995b;a; Douc et al., 2004; Meyn & Tweedie, 1994; Baxendale, 2005). Our last assumption formalizes the requirement of a finite moment for the gradient error.

**Assumption 5.** *There is $q > 1$ such that for $\widetilde{G}(\theta,\zeta) \sim \mathcal{D}_{\mathcal{I}}(\theta)$, we have*

$$\mathbb{E}\big[\|\varepsilon_\theta\|^q \mid \theta\big]^{1/q} = \mathbb{E}\big[\|\widetilde{G}(\theta,\zeta) - \nabla\mathcal{L}(\theta)\|^q \mid \theta\big]^{1/q} \le A_q\|\theta-\theta^\star\| + B_q \tag{4}$$

*for all $\theta \in \mathbb{R}^d$, where $A_q, B_q > 0$. When $\mathcal{L}$ is not strongly convex, we further assume that $A_q = 0$.*

The bound (4) captures the case of arbitrarily high noise magnitude through the dependence on $\|\theta - \theta^\star\|$. This is consistent with convex optimization problems with $L$-Lipschitz-smooth objectives (Assumption 1) where the norm of the gradient $\|\nabla\mathcal{L}(\theta)\|$ is bounded by $L \cdot \|\theta - \theta^\star\|$. Assumption 5 improves upon the conditions used in (Gorbunov et al., 2020; Tsai et al., 2022; Gorbunov et al., 2023; Nguyen et al., 2023) since these either required a uniformly constant upperbound (independent of $\theta$) or only considered the case $q = 2$ (finite variance). For non-strongly convex $\mathcal{L}$, we require $A_q = 0$ since $\theta^\star$ may not exist.

**Definition 1.** *If $X$ is a real random variable, we say that $X$ is $K$-sub-Gaussian for $K > 0$ if*

$$\mathbb{E}\exp(\lambda^2X^2) \le e^{\lambda^2K^2} \quad \textit{for} \quad |\lambda| \le 1/K. \tag{5}$$

*We say that $X$ is $K$-sub-exponential for $K > 0$ if*

$$\mathbb{E}\exp(\lambda|X|) \le \exp(\lambda K) \quad \textit{for all} \quad 0 \le \lambda \le 1/K. \tag{6}$$

The convergence results presented in this paper use the following formalism of continuous state Markov chains. Given a step size $\beta > 0$ and a quantile $p \in (0,1)$, we denote by $P_{\beta,p}$ the Markov transition kernel governing the Markov chain $(\theta_t)_{t\ge0}$ generated by QC-SGD, so that

$$\mathbb{P}(\theta_{t+1} \in A \mid \theta_t) = P_{\beta,p}(\theta_t, A)$$

for $t \ge 0$ and $A \in \mathcal{B}(\mathbb{R}^d)$. The transition kernel $P_{\beta,p}$ acts on probability distributions $\nu \in \mathcal{M}_1(\mathbb{R}^d)$ through the mapping $\nu \to \nu P_{\beta,p}$ which is defined, for all $A \in \mathcal{B}(\mathbb{R}^d)$, by $\nu P_{\beta,p}(A) = \int_A P_{\beta,p}(\theta, A)d\nu(\theta) = \mathbb{P}(\theta_{t+1} \in A|\theta_t \sim \nu)$. For $n \ge 1$, we similarly define the multi-step transition kernel $P^n_{\beta,p}$ which is such that $P^n_{\beta,p}(\theta_t, A) =$

$\mathbb{P}(\theta_{t+n} \in A \mid \theta_t)$ and acts on probability distributions $\nu \in \mathcal{M}_1(\mathbb{R}^d)$ through $\nu P^n_{\beta,p} = (\nu P_{\beta,p})P^{n-1}_{\beta,p}$. Finally, we define the total variation (TV) norm of a signed measure $\nu$ as

$$2\|\nu\|_{\mathrm{TV}} = \sup_{f:|f|\leq 1} \int f(\theta)\nu(d\theta) = \sup_{A\in\mathcal{B}(\mathbb{R}^d)} \nu(A) - \inf_{A\in\mathcal{B}(\mathbb{R}^d)} \nu(A). \tag{7}$$

In particular, we recover the TV *distance* between $\nu_1, \nu_2 \in \mathcal{M}_1(\mathbb{R}^d)$ as

$$d_{\mathrm{TV}}(\nu_1, \nu_2) = \|\nu_1 - \nu_2\|_{\mathrm{TV}} = \sup_{A\in\mathcal{B}(\mathbb{R}^d)} \big|\nu_1(A) - \nu_2(A)\big|.$$

The second equality reflects the fact that the TV distance between two probability measures corresponds to the largest absolute difference between the probabilities they assign to the same event. The TV distance is a broadly used metric to quantify the convergence of Markov chains (Levin & Peres, 2017; Baxendale, 2005; Meyn & Tweedie, 1993; Rosenthal, 1995a) besides the Wasserstein distance (Dieuleveut et al., 2020).

In the next section, we will prove that the Markov chain defined by iteration (2) converges to unique invariant distribution in TV distance. This convergence mode will allow us to extrapolate the properties of the limit distribution on the iterates $\theta_t$ and thus derive non-asymptotic concentration bounds for them, see Corollaries 2 and 1 below.

## 3 Strongly Convex Objectives

We are ready to state our convergence result for the stochastic optimization of a strongly convex objective using QC-SGD with $\eta$-corrupted samples.

**Theorem 1** (Geometric ergodicity)**.** *Let Assumptions 1-5 hold and assume there is a quantile $p \in [\eta, 1-\eta]$ such that*

$$\kappa := (1-\eta)p\mu - \eta L - (1-p)^{-\frac{1}{q}} A_q(1 - p(1-\eta)) > 0. \tag{8}$$

*Then, for a step size $\beta$ satisfying*

$$\beta < \frac{1}{4}\frac{\kappa}{\mu^2 + 24\eta L^2 + 28A_q^2} \wedge \frac{2}{\mu + L}, \tag{9}$$

*the Markov chain $(\theta_t)_{t\geq 0}$ generated by QC-SGD with parameters $\beta$ and $p$ converges geometrically to a unique invariant measure $\pi_{\beta,p}$: for any initial $\theta_0 \in \mathbb{R}^d$, there is $\rho < 1$ and $M < \infty$ such that after $T$ iterations*

$$\big\|\delta_{\theta_0} P^T_{\beta,p} - \pi_{\beta,p}\big\|_{\mathrm{TV}} \leq M\rho^T\big(1 + \|\theta_0 - \theta^\star\|^2\big),$$

*where $\delta_{\theta_0}$ is the Dirac measure located at $\theta_0$.*

The proof of Theorem 1 is given in Appendix D.3 and relies on the geometric ergodicity result of (Meyn & Tweedie, 1993, Chapter 15) for Markov chains with a geometric drift property. A similar result for quadratically growing objectives was established by (Yu et al., 2021) and convergence w.r.t. Wasserstein's metric was shown in (Dieuleveut et al., 2020) assuming gradient co-coercivity. However, robustness was not considered in these works. Theorem 1 establishes the iteration's convergence to a unique invariant measure $\pi_{\beta,p}$. The properties of this limit distribution will be explored in the sequel. The restriction $p \in [\eta, 1-\eta]$ comes from the consideration that other quantiles are not estimable in the event of $\eta$-corruption. Condition (8) is best interpreted for the choice $p = 1-\eta$ in which case it translates into $\eta^{1-1/q} \leq \mathcal{O}(\mu/(L + A_q))$ implying that it is verified for $\eta$ small enough within a limit fixed by the problem conditioning. A similar condition with $q = 2$ appears in (Diakonikolas et al., 2022, Theorem E.9) which uses a finite variance assumption.

When (8) is satisfied, one clearly has that $\kappa = \mathcal{O}(\mu)$. Considering $q = 2$ for simplicity and taking the maximum allowed corruption rate in this case $\eta = \mathcal{O}(\mu^2/(L + A_q)^2)$ leads to an upperbound on the step-size $\beta$ of order $\mathcal{O}(\mu/(\mu^2 + A_q^2) \wedge 1/L)$. While the condition $\beta = \mathcal{O}(1/L)$ is standard in smooth optimization, the additional condition in terms of $A_q$ ensures that the noise introduced to the iteration by the gradient samples does not cause it to diverge.

The constants $M$ and $\rho$ controlling the geometric convergence speed in Theorem 1 depend on the parameters $\beta, p$ and the initial $\theta_0$. Among choices fulfilling the convergence conditions, it is straightforward that greater step size $\beta$ and $\theta_0$ closer to $\theta^\star$ lead to faster convergence. However, the dependence in $p$ is more intricate and should be evaluated through the resulting value of $\kappa$. We provide a more detailed discussion about the value of $\rho$ in Appendix C.

The choice $p = 1 - \eta$ appears to be ideal since it leads to optimal deviation of the invariant distribution around the optimum $\theta^\star$ which is the essence of our next statement.

**Proposition 1.** *Assume the same as in Theorem 1 and condition (8) with the choice $p = 1 - \eta$. For step size $\beta$ satisfying (9), $q \geq 2$, and additionally:*

$$\beta \leq \eta^{2-2/q}/\kappa, \tag{10}$$

*for $\theta \sim \pi_{\beta,1-\eta}$, we have the following upper bound:*

$$\mathbb{E}\|\theta - \theta^\star\|^2 \leq \Big(\frac{6\eta^{1-1/q}B_q}{\kappa}\Big)^2.$$

Proposition 1 is proven in Appendix D.4. An analogous result holds for $q \in (1, 2)$ but requires a different proof and can be found in Appendix D.5. Proposition 1 may be compared to (Yu et al., 2021, Theorem 3.1) which shows that the asymptotic estimation error can be reduced arbitrarily using a small step size. However, this is impossible in our case since we consider corrupted gradients. The performance of Proposition 1 is best discussed in the specific context of linear regression where gradients are given as $G(\theta, (X, Y)) = X(X^\top\theta - Y)$ for samples $X, Y \in \mathbb{R}^d \times \mathbb{R}$ such that $Y = X^\top\theta^\star + \epsilon$ with $\epsilon$ a centered noise. In this case, a finite moment of order $k$ for the data implies order $k/2$ for the gradient corresponding to an $\eta^{1-2/k}$ rate in Proposition 1. Since Assumption 5 does not include independence of the noise $\epsilon$ from $X$, this corresponds to the negatively correlated moments assumption of (Bakshi & Prasad, 2021) being unsatisfied. Consequently, Proposition 1 is information-theoretically optimal in $\eta$ based on (Bakshi & Prasad, 2021, Corollary 4.2). Nonetheless, the dimension dependence through $B_q$ remains poor since we have $B_q \sim \sqrt{d}$ in general because the Euclidean norm is used in Assumption 5. This dimension dependence may be improvable by using the quantiles $\sup_{\|v\|=1} Q_p\big(|\langle \widetilde{G}(\theta_t, \zeta_t), v\rangle|\big)$ as clipping thresholds and adapting ideas from (Catoni & Giulini, 2018) in the analysis. However, exploring this method is beyond our scope as the involved estimations for all $\|v\| = 1$ would be excessively sample hungry and computationally heavy for stochastic optimization.If the gradient is sub-Gaussian with constant $K$, we would have $B_q \lesssim K\sqrt{q}$ for $q \geq 1$ (see (Vershynin, 2018) for a reference), in which case, the choice $q = \log(1/\eta)$ recovers the optimal rate in $\eta\sqrt{\log(1/\eta)}$ for the Gaussian case.

We now turn to showing strong concentration properties for the invariant distribution $\pi_{\beta,p}$. For this purpose, we restrict the optimization to a bounded and convex set $\Theta \subset \mathbb{R}^d$ and replace Iteration (2) by the projected iteration

$$\theta_{t+1} = \Pi_\Theta\big(\theta_t - \alpha_{\theta_t}\beta G(\theta_t, \zeta_t)\big), \tag{11}$$

where $\Pi_\Theta$ is the projection onto $\Theta$. Assuming that the latter contains the optimum $\theta^\star \in \Theta$, one can check that the previous results continue to hold thanks to the inequality

$$\|\Pi_\Theta(\theta) - \theta^\star\| = \|\Pi_\Theta(\theta) - \Pi_\Theta(\theta^\star)\| \leq \|\theta - \theta^\star\|,$$

which results from the convexity of $\Theta$. The restriction of the optimization to a bounded set allows us to uniformly bound the clipping threshold $\tau_\theta$, which is indispensable for the following result.

**Proposition 2.** *In the setting of Theorem 1, consider projected QC-SGD (11) and let $\overline{\tau} = \sup_{\theta\in\Theta}\tau_\theta, D = \mathrm{diam}(\Theta)$ the diameter of $\Theta$ and $\overline{B}_q = A_qD + B_q$.*

- *Consider the non-corrupted case $\eta = 0$ and set the quantile $p$ such that $p \geq 1 - (\beta\mu)^{\frac{q}{2(q-1)}}$. Then, for $\theta \sim \pi_{\beta,p}$, the variable $\|\theta - \theta^\star\|$ is sub-Gaussian in the sense of Definition 1 with constant*

$$K = 4\sqrt{\frac{2\beta(\overline{B}_q^2 + \overline{\tau}^2)}{p\mu}}.$$

- *Consider the corrupted case $\eta > 0$, and set the quantile $p \in [\eta, 1-\eta]$ such that Inequality (8) holds. Then, for $\theta \sim \pi_{\beta,p}$, the variable $\|\theta - \theta^\star\|$ is sub-exponential in the sense of Definition 1 with constant*

$$K = \frac{7\overline{\tau} + (1-p)^{1-1/q}\overline{B}_q}{p\mu}.$$

The proof can be found in Appendix D.6. The strong concentration properties given by Proposition 2 for the invariant distribution appear to be new. Still, the previous result remains asymptotic in nature. High confidence deviation bounds for an iterate $\theta_t$ can be derived by leveraging the convergence in Total Variation distance given by Theorem 1 leading to the following result.

**Corollary 1.** *In the setting of Proposition 2, in the absence of corruption $\eta = 0$, after $T$ iterations, for $\delta > 0$, we have*

$$\mathbb{P}\Big( \big\|\theta_T - \theta^\star\big\| > 4\sqrt{\overline{B}_q^2 + \overline{\tau}^2}\sqrt{\frac{2\beta\log(e/\delta)}{p\mu}} \Big) \leq \delta + \rho^T M\big(1 + \|\theta_0 - \theta^\star\|^2\big).$$

Choosing a smaller step size $\beta$ in Corollary 1 allows to improve the deviation bound. However, this comes at the cost of weaker confidence because of slower convergence due to a greater $\rho$. See Appendix C for a discussion including a possible compromise. Corollary 1 may be compared to the results of (Gorbunov et al., 2020; Tsai et al., 2022; Sadiev et al., 2023; Nguyen et al., 2023) which correspond to $\beta \approx 1/T$ and have a similar dependence on the dimension through the gradient variance. Although their approach is also based on gradient clipping, they use different thresholds and proof methods. In the presence of corruption, the invariant distribution is not sub-Gaussian. This can be seen by considering the following toy Markov chain:

$$X_{t+1} = \begin{cases} \alpha X_t + \xi & \text{w.p.} \quad 1-\eta \\ X_t + \tau & \text{w.p.} \quad \eta \end{cases}$$

where $\alpha < 1, \tau > 0$ are constants and $\xi$ is a positive random noise. Using similar methods to the proof of Theorem 1, one can show that $(X_t)_{t\geq 0}$ converges (for any initial $X_0$) to an invariant distribution whose moments can be shown to grow linearly, indicating a sub-exponential distribution and excluding a sub-Gaussian one. We provide additional details for the underlying argument in Appendix D.7. For the corrupted case, the sub-exponential property stated in Proposition 2 holds with a constant $K$ of order $\overline{\tau}/\mu$, which is not satisfactory and leaves little room for improvement due to the inevitable bias introduced by corruption. Therefore, we propose the following procedure in order to obtain a high confidence estimate, similarly to Corollary 1.

**Algorithm 1:** Aggregation of cycling iterates

**Input:** Step size $\beta > 0$, quantile index $p \in (0,1)$, initial parameter $\theta_0 \in \Theta$, horizon $T$ and number of concurrent iterates $N \geq 1$.

Optimize multiple parameters $\theta_t^{(1)}, \ldots, \theta_t^{(N)}$ starting from a common $\theta_0 = \theta_0^{(n)}$ for $n \in [\![N]\!] =: \{1, \ldots, N\}$ and $T$ steps $t = 0, \ldots, T$ using the following cycling iteration:

$$\theta_{t+1}^{(n)} = \begin{cases} \theta_t^{(n)} - \alpha_{\theta_t^{(n)}}\beta G\big(\theta_t^{(n)}, \zeta_t\big) & \text{if } t \equiv n-1 \bmod N, \\ \theta_t^{(n)} & \text{otherwise.} \end{cases} \tag{12}$$

Compute the pairwise distances $r_{i,j} = \big\|\theta_T^{(i)} - \theta_T^{(j)}\big\|$ for $i, j \in [\![N]\!]$.

For $i \in [\![N]\!]$, let $r^{(i)} \in \mathbb{R}_+^N$ be the vector $r_{i,:} := [r_{i,1}, \ldots, r_{i,N}]$ sorted in non decreasing order.

Compute the aggregated estimator as $\widehat{\theta} = \theta_T^{(\widehat{i})}$ with $\widehat{i} = \arg\min_{i \in [\![N]\!]} r^{(i)}_{\lfloor N/2 \rfloor}$.

**return** $\widehat{\theta}$

Algorithm 1 uses ideas from (Hsu & Sabato, 2016) (see also (Minsker, 2015; Juditsky et al., 2023)) and combines the collection of *weak* estimators $\big(\theta_T^{(i)}\big)_{i\in[\![N]\!]}$ (only satisfying $L_2$ bounds) into a strong one with sub-exponential deviation. This is done by picking $\theta_T^{(i)}$ which is such that the median of its distances to other estimators $r_{\lfloor N/2 \rfloor}^{(i)}$ is minimal. The aggregated estimator $\widehat{\theta}$ satisfies the high probability bound given in the next result.

**Corollary 2.** *Assume the same as in Theorem 1 and Proposition 1. Consider $\widehat{\theta}$ given by Algorithm 1, with the assumption that the gradient sample sets used for each $\big(\theta_T^{(n)}\big)_{n\in[\![N]\!]}$ in Equation (12) are independent. For $\delta > 0$, if $N \geq 16\log(1/\delta)$ and $T$ satisfies*

$$T \geq N\log(15M(1+\|\theta_0-\theta^\star\|^2))/\log(1/\rho),$$

*then, with probability at least $1-\delta$, we have*

$$\big\|\widehat{\theta}-\theta^\star\big\| \leq \frac{27\eta^{1-\frac{1}{q}}\overline{B}_q}{\kappa}. \tag{13}$$

We obtain a high confidence version of the bound in expectation previously stated in Proposition 1. As argued before, the above bound depends optimally on $\eta$. Similar bounds to (13) are obtained for $q = 2$ in (Diakonikolas et al., 2022) for streaming mean estimation, linear and logistic regression. Their results enjoy better dimension dependence but are less general than ours since we handle the case $q \in (1, 2)$ and consider strongly convex objectives more broadly. In addition, our results further extend to non-convex objectives as detailed in the next section. Finally, the implementation of the algorithm in (Diakonikolas et al., 2022) is not straightforward whereas our method is quite easy to use (see Section 5).

## 4 Non-Convex Objectives

In this section, we drop Assumption 2 and consider the optimization of possibly non-convex objectives. Consequently, the existence of a unique optimum $\theta^\star$ and the quadratic growth of the objective are no longer guaranteed. This motivates us to use a uniform version of Assumption 5 with $A_q = 0$ since the gradient is no longer assumed coercive and its deviation moments can be taken as bounded. In this context, we obtain the following weaker (compared to Theorem 1) ergodicity result for QC-SGD.

**Theorem 2** (Ergodicity)**.** *Let Assumptions 1, 3, 4 and 5 hold with $A_q = 0$ (uniformly bounded moments) and let $\mathcal{L}$ be an objective such that $\inf_\theta \mathcal{L}(\theta) > -\infty$ is finite. Let $(\theta_t)_{t\geq 0}$ be the Markov chain generated by QC-SGD with step size $\beta$ and quantile $p \in [\eta, 1-\eta]$. Assume that $p$ and $\beta$ are such that $3p(1-\eta)/4 > L\beta+\eta$ and that the subset of $\mathbb{R}^d$ given by*

$$\Big\{\theta : \frac{1}{2}\big\|\nabla\mathcal{L}(\theta)\big\|^2 \leq \frac{B_q^2\big((1-p)^{-\frac{2}{q}}(L\beta+2\eta^2)+2\eta^{2-\frac{2}{q}}\big)}{p(1-\eta)(3p(1-\eta)/4-L\beta-\eta)}\Big\} \tag{14}$$

*is bounded. Then, for any initial $\theta_0 \in \mathbb{R}^d$, there exists $M < +\infty$ such that after $T$ iterations*

$$\big\|\delta_{\theta_0}P_{\beta,p}^T - \pi_{\beta,p}\big\|_{\mathrm{TV}} \leq \frac{M}{T}, \tag{15}$$

*where $\pi_{\beta,p}$ is a unique invariant measure and where $\delta_{\theta_0}$ is the Dirac measure located at $\theta_0$.*

The proof is given in Appendix D.10 and uses ergodicity results from (Meyn & Tweedie, 1993, Chapter 13). Theorem 2 provides convergence conditions for an SGD Markov chain on a smooth objective in a robust setting. We are unaware of anterior results of this kind in the literature. Condition (14) requires that the set where the true gradient norm is smaller than the estimation error is bounded. This aims to exclude the possibility that the iteration gets trapped within this set and keep using unreliable gradient estimates causing it to diverge. The result is stronger when the upperbound in (14) is smaller. Note that setting $p$ close to $1-\eta$ increases the clipping threshold and the estimation error as a consequence, making this condition harder to satisfy. On the other hand, using $\beta = \mathcal{O}(1/L)$ and a more conservative value of $p$ makes the upperbound of order $\mathcal{O}(B_q^2)$ and condition 14 easier to satisfy. Observe that, for no corruption ($\eta = 0$), the condition is always fulfilled for some $\beta$ and $p$. Note also that without strong convexity (Assumption 2), convergence occurs at a slower sublinear rate which is consistent with the optimization rate expected for a smooth objective (see (Bubeck, 2015, Theorem 3.3)).

As previously, we complement Theorem 1 with a characterization of the invariant distribution.

**Proposition 3.** *Under the conditions of Theorem 2, assume that the choice $p = 1 - \eta$ is such that the set (14) is bounded. For step size $\beta \leq \eta^2/L$, the stationary measure $\theta \sim \pi_{\beta,1-\eta}$ satisfies*

$$\mathbb{E}\big\|\nabla\mathcal{L}(\theta)\big\|^2 \leq \frac{5\eta^{2-\frac{2}{q}}B_q^2}{p(1-\eta)\big(3p(1-\eta)/4 - L\beta - \eta\big)}. \tag{16}$$

The statement of Proposition 3 is clearly less informative than Propositions 1 and 2 since it only pertains to the gradient rather than, for example, the excess risk. This is due to the weaker assumptions that do not allow to relate these quantities. Still, the purpose remains to find a critical point and is achieved up to $\mathcal{O}(\eta^{1-1/q})$ precision according to this result. Due to corruption, the estimation error on the gradient cannot be reduced beyond $\Omega(\eta^{1-1/q})$ (Prasad et al., 2020; Hopkins & Li, 2018; Diakonikolas & Kane, 2019). Therefore, one may draw a parallel with a corrupted mean estimation task, in which case, the previous rate is, in fact, information-theoretically optimal.

## 5 Implementation and Numerical Experiments

The use of the generally unknown quantile $Q_p(\|\widetilde{G}(\theta_t, \zeta_t)\|)$ in QC-SGD constitutes the main obstacle to its implementation. For strongly convex objectives, one may use a proxy such as $a\|\theta_t - \theta_{\text{ref}}\| + b$ with positive $a, b$ and $\theta_{\text{ref}} \in \mathbb{R}^d$ an approximation of $\theta^\star$ serving as reference point. This choice is consistent with Assumptions 1 and 5, see Lemma 2 in Appendix D. For instances of Problem (1) defined with an

**Algorithm 2:** Rolling QC-SGD

**Input:** Step size $\beta > 0$, quantile index $p \in (0, 1)$, initial parameter $\theta_0 \in \mathbb{R}^d$, $\tau_{\text{unif}} > 0$, buffer $B$ of size $S$ and horizon $T$.

Fill $B$ with $S - 1$ values equal to $\tau_{\text{unif}}$.
**for** $t = 0 \dots T - 1$ **do**
  Draw a sample $G(\theta_t, \zeta_t)$ and add $\|G(\theta_t, \zeta_t)\|$ to $B$.
  $\widehat{Q}_p \leftarrow \lfloor pS \rfloor$ rank element of $B$.
  $\theta_{t+1} \leftarrow \theta_t - \beta\text{clip}(G(\theta_t, \zeta_t), \widehat{Q}_p)$
  Delete the oldest value in $B$.
**return** $\theta_T$

asymptotically linear function $\ell$ such as the logistic, hinge or Huber's loss, a constant threshold can be used since the gradient is a priori uniformly bounded, implying the same for the quantiles of its deviations. In practice, we propose a simpler and more direct approach: we use a rolling quantile procedure, described in Algorithm 2. The latter stores the values $(\|G(\theta_{t-j}, \zeta_{t-j})\|)_{1\leq j\leq S}$ in a buffer of size $S \in \mathbb{N}^*$ and replaces $Q_p(\|\widetilde{G}(\theta_t, \zeta_t)\|)$ in QC-SGD by an estimate $\widehat{Q}_p$ which is the $\lfloor pS \rfloor$-th order statistic in the buffer. Note that only the norms of previous gradients are stored in the buffer, limiting the memory overhead to $\mathcal{O}(S)$. The computational cost of $\widehat{Q}_p$ can also be kept to $\mathcal{O}(S)$ per iteration thanks to a bookkeeping procedure (see Appendix B).

Note that, since Algorithm 2 uses the corrupted samples $G(\theta_t, \zeta_t)$ rather than the true ones $\widetilde{G}(\theta_t, \zeta_t)$ to estimate the quantiles, a more conservative upperbound of roughly $p \leq 1 - 2\eta$ should be respected when an estimate of $\eta$ is available. Otherwise, one may default to $p = 1/2$ as an initial guess and adapt based on performance. In practice, our experiments show that relatively low values within $p \in [0.1, 0.2]$ are best for strongly convex objectives while higher values are affordable in other cases. See Appendix B for more details.

We implement this procedure for a few tasks and compare its performance with relevant baselines. We do not include a comparison with (Diakonikolas et al., 2022) whose procedure has no implementation we are aware of and is difficult to use in practice. Indeed, the algorithm in question heavily depends on several problem parameters and involves a filtering procedure which requires multiple passes on large data mini-batches making it impractical for the streaming setting. Moreover, a number of special methods are required to mitigate the costs of the matrix operations needed in the original procedure making the algorithm's implementation even more involved.

Our experiments on synthetic data consider an infinite horizon, dimension $d = 128$, and a constant step size for all methods.

**Linear regression.** We consider least-squares linear regression and compare RQC-SGD with Huber's estimator (Huber, 1973) and clipped SGD (designated as CClip($\lambda$)) with three clipping levels $\lambda\sigma_{\max}\sqrt{d}$ for $\lambda \in \{0.8, 1.0, 1.2\}$ where $\sigma_{\max}$ is a fixed data scaling factor. These thresholds provide a rough estimate of the gradient norm near the optimum $\theta^\star$. We generate covariates $X$ and labels $Y$ both heavy-tailed and corrupted. Corruption in the data stream is generated according to Assumption 3 with outliers represented either by aberrant values or *fake* samples $Y = X^\top\theta_{\text{fake}} + \epsilon$ using a false parameter $\theta_{\text{fake}}$, see Appendix B for further details on data generation and fine tuning of the Huber parameter. All methods are run with constant step size and averaged results over 100 runs are displayed on Figure 1 (top row).

As anticipated, Huber's loss function is not robust to corrupted covariates. In contrast, using gradient clipping allows convergence to meaningful estimates. Although this holds true for a constant threshold, Figure 1 shows it may considerably slow the convergence if started away from the optimum. In addition, the clipping level also affects the final estimation precision and requires tuning. Both of the previous issues are well addressed by RQC-SGD whose adaptive clipping level allows fast progress of the optimization and accurate convergence towards a small neighborhood of the optimum.

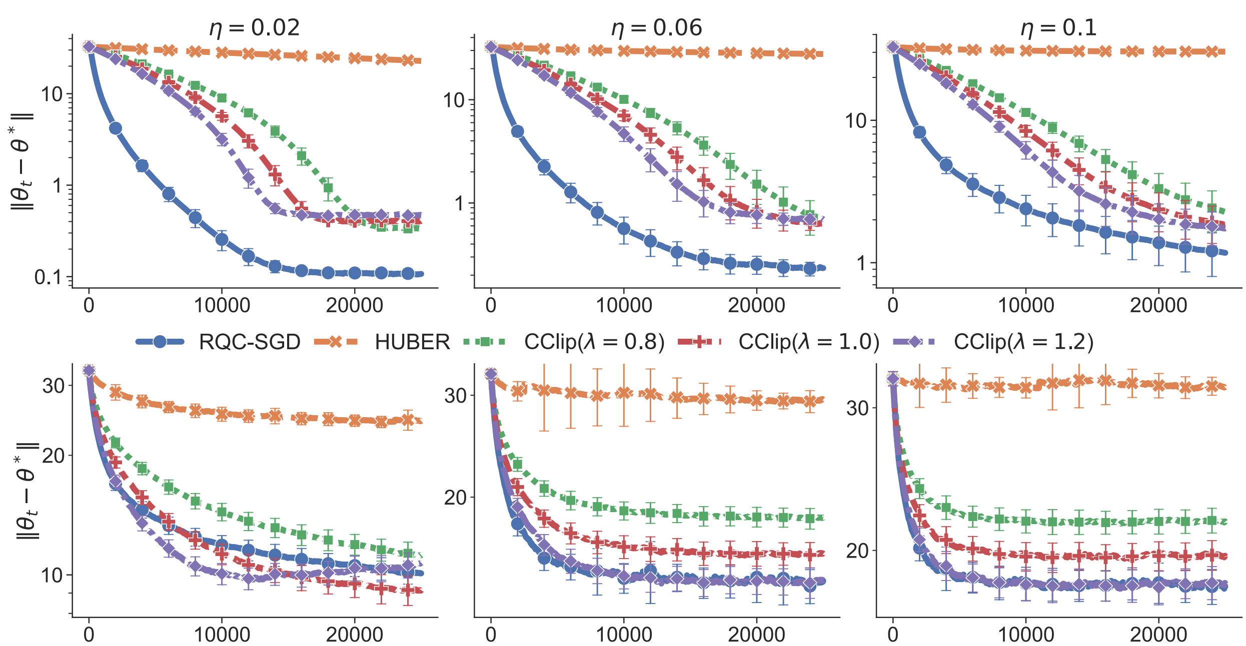


Figure 1: Evolution of $\|\theta_t - \theta^\star\|$ on the tasks of linear regression (top row) and logistic regression (bottom row) averaged over 100 runs at increasing corruption levels (error bars represent half the standard deviation). Estimators based on Huber's loss are strongly affected by data corruption. SGD with constant clipping thresholds is robust but slow to converge for linear regression and requires tuning for better final precision. RQC-SGD combines fast convergence with good final precision thanks to its adaptive clipping strategy.

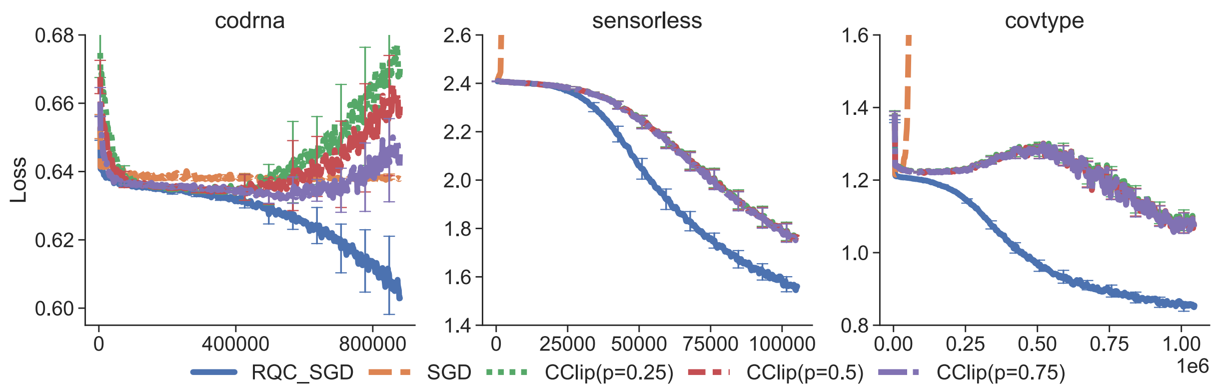


Figure 2: Evolution of the test loss ($y$-axis) against iteration $t$ ($x$-axis) for the training of a single hidden layer network on different real world classification datasets (average over 20 runs). We observe more consistent and stable objective decrease for RQC-SGD whereas constant clipping baselines are slower and may fail to converge.

**Logistic regression.** We test the same methods on logistic regression. Huber's baseline is represented by the modified Huber loss (also known as quadratic SVM (Zhang, 2004)). We generate data similarly to the previous task except for the labels which follow $Y \sim \text{Bernoulli}(\sigma(X^\top \theta^\star))$ with $\sigma$ the sigmoid function. Corrupted labels are either uninformative, flipped or obtained with a fake $\theta_{\text{fake}}$ (see details in Appendix B). Results are displayed on the bottom row of Figure 1.

As previously, Huber's estimator performs poorly with corruption. However, constant clipping appears to be better suited when the gradient is bounded, so that the optimization is less affected by its underestimation. We observe, nonetheless, that a higher clipping level may lead to poor convergence properties, even at a low corruption rate. Note also that the constant levels we use are based on prior knowledge about the data distribution and would have to be fine tuned in practice. Meanwhile, the latter issue is well addressed by quantile clipping. Finally, notice that no algorithm truly approaches the true solution for this task. This reflects the difficulty of improving upon Proposition 3 which only states convergence to a neighborhood where the objective gradient is comparable to the estimation error in magnitude.

**Classification with shallow networks.** Finally, we evaluate the performance on the task of training a single hidden layer neural network classifier on some real datasets which corresponds to a non-convex optimization problem. To handle multiclass data, we use the cross entropy loss and replace Huber's baseline with plain SGD for simplicity. We define constant clipping baselines using thresholds given by the quantiles of order $p = 0.25, 0.5$, and $0.75$ of the norms of a batch of gradients at the beginning of the optimisation. Due to the greater sensitivity to corruption observed in this case, we set $\eta = 0.02$ and use $p = 0.9$ for RQC-SGD. We train all methods with one sample per iteration using equal step sizes and evaluate them through the test loss. We provide further results and experimental details in Appendix B. Results are displayed on Figure 2.

Unsurprisingly, standard SGD is not robust to corrupted samples and, while using a constant clipping level helps keep the optimisation on track, the experiments show that careful tuning may sometimes be necessary to prevent divergence. On the other hand, the adaptive clipping levels used by RQC-SGD allow to make the iteration faster *and* more resilient to corruption. This leads to an optimization path with a more consistent decrease of the objective. Moreover, we also observe that RQC-SGD allows for a better control of the asymptotic variance of the optimized parameter compared to constant clipping.

## 6 Conclusion

We introduced a new clipping strategy for SGD and proved that it defines a stochastic optimization procedure which is robust to both heavy tails and outliers in the data stream. We also provided an efficient rolling quantile procedure to implement it and demonstrated its performance through numerical experiments on synthetic and real data. Future research directions include improving the dimension dependence in our bounds, possibly by using sample rejection rules or by considering stochastic mirror descent (Nemirovskij & Yudin, 1983; Beck & Teboulle, 2003) clipped with respect to a non Euclidean norm. This may also procure robustness to higher corruption rates. Another interesting research track is the precise quantification of the geometric convergence speed of the Markov chain generated by constant step size SGD on a strongly convex objective.

### Broader Impact Statement

In this optional section, TMLR encourages authors to discuss possible repercussions of their work, notably any potential negative impact that a user of this research should be aware of. Authors should consult the TMLR Ethics Guidelines available on the TMLR website for guidance on how to approach this subject.

### Author Contributions

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors. Only add this information once your submission is accepted and deanonymized.

### Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper. Only add this information once your submission is accepted and deanonymized.

## References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, pp. 20–29, 1996.

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 102–115, 2021.

Martyna Bator. Dataset for Sensorless Drive Diagnosis. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5VP5F.

Peter H Baxendale. Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *The Annals of Applied Probability*, 15(1B):700–738, 2005.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Witold Bednorz. The Kendall's Theorem and its Application to the Geometric Ergodicity of Markov Chains. *arXiv preprint arXiv:1301.1481*, 2013.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*, volume 22. Springer Science & Business Media, 2012.

Kenneth S Berenhaut and Robert Lund. Geometric renewal convergence rates from hazard rates. *Journal of Applied Probability*, 38(1):180–194, 2001.

Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50K5N.

Joseph K Blitzstein and Jessica Hwang. *Introduction to probability.* Crc Press, 2019.

D. A. Bloch. A note on the estimation of the location parameter of the cauchy distribution. *Journal of the American Statistical Association*, 61:852–855, 1966.

Léon Bottou and Yann Cun. Large scale online learning. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/9fb7b048c96d44a0337f049e0a61ff06-Paper.pdf.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Léon Bottou and Yann Lecun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21:137 – 151, 03 2005. doi: 10.1002/asmb.538.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18 – 43, 2013. doi: 10.3150/11-BEJ390. URL https://doi.org/10.3150/11-BEJ390.

Hervé Cardot, Peggy Cénac, and Antoine Godichon-Baggioni. Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, 45(2):591 – 614, 2017. doi: 10.1214/16-AOS1460. URL https://doi.org/10.1214/16-AOS1460.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148 – 1185, 2012. doi: 10.1214/11-AIHP454. URL https://doi.org/10.1214/11-AIHP454.

Olivier Catoni and Ilaria Giulini. Dimension-free pac-bayesian bounds for the estimation of the mean of a random vector. *arXiv preprint arXiv:1802.04308*, 2018.

Han-fu Chen and AI-JUN Gao. Robustness analysis for stochastic approximation algorithms. *Stochastics and Stochastic Reports*, 26(1):3–20, 1989.

Han-Fu Chen, Lei Guo, and Ai-Jun Gao. Convergence and robustness of the robbins-monro algorithm truncated at randomly varying bounds. *Stochastic Processes and their Applications*, 27:217–231, 1987.

Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pp. 786–806. PMLR, 2019.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27, 2014.

Jules Depersin and Guillaume Lecué. Robust sub-Gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1):511–536, 2022. doi: 10.1214/21-AOS2118. URL https://doi.org/10.1214/21-AOS2118.

Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

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

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.

Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33:1830–1840, 2020.

Ilias Diakonikolas, Daniel M Kane, Ankit Pensia, and Thanasis Pittas. Streaming algorithms for high-dimensional robust statistics. In *International Conference on Machine Learning*, pp. 5061–5117. PMLR, 2022.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348 – 1382, 2020. doi: 10.1214/19-AOS1850. URL https://doi.org/10.1214/19-AOS1850.

R. Douc, E. Moulines, and Jeffrey S. Rosenthal. Quantitative bounds on convergence of time-inhomogeneous Markov chains. *The Annals of Applied Probability*, 14(4):1643 – 1665, 2004. doi: 10.1214/105051604000000620. URL https://doi.org/10.1214/105051604000000620.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. *arXiv preprint arXiv:2310.01860*, 2023.

Frank R Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42 (6):1887–1896, 1971.

Abdelhakim Hannousse and Salima Yahiouche. Web page phishing detection. *Mendeley Data*, 2, 2020.

Matthew Holland and Kazushi Ikeda. Better generalization with less data using robust gradient descent. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2761–2770. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/holland19a.html.

Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, 2020. doi: 10.1214/19-AOS1843. URL https://doi.org/10.1214/19-AOS1843.

Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1021–1034, 2018.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pp. 1753–1758, 1965.

Peter J Huber. The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, 43(4):1041–1067, 1972.

Peter J. Huber. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799 – 821, 1973. doi: 10.1214/aos/1176342503. URL https://doi.org/10.1214/aos/1176342503.

Peter J Huber. Robust estimation of a location parameter. *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518, 1992.

Daniel C Jerison. Quantitative convergence rates for reversible Markov chains via strong random times. *arXiv preprint arXiv:1908.06459*, 2019.

Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

Anatoli Juditsky, Andrei Kulunchakov, and Hlib Tsyntseus. Sparse recovery by reduced variance stochastic approximation. *Information and Inference: A Journal of the IMA*, 12(2):851–896, 2023.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer New York, NY, 2003.

Guanghui Lan. *First-Order and Stochastic Optimization Methods for Machine Learning*, volume 1. Springer, 2020.

Guillaume Lecué and Matthieu Lerasle. Learning from mom's principles: Le cam's approach. *Stochastic Processes and their applications*, 129(11):4385–4410, 2019.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48, 11 2017. doi: 10.1214/19-AOS1828.

Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. A fast spectral algorithm for mean estimation with sub-Gaussian rates. In *Conference on Learning Theory*, pp. 2598–2612. PMLR, 2020.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 411–421. PMLR, 2020.

Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393 – 410, 2021. doi: 10.1214/20-AOS1961. URL `https://doi.org/10.1214/20-AOS1961`.

Robert B Lund, Sean P Meyn, and Richard L Tweedie. Computable exponential convergence rates for stochastically ordered Markov processes. *The Annals of Applied Probability*, 6(1):218–237, 1996.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

Rainer Martin and Carl Masreliez. Robust estimation via stochastic approximation. *IEEE Transactions on Information Theory*, 21(3):263–271, 1975. doi: 10.1109/TIT.1975.1055386.

H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1222–1230, 2013.

Sean P Meyn and Richard L Tweedie. *Markov Chains and Stochastic Stability.* Springer London, 1993.

Sean P Meyn and Robert L Tweedie. Computable bounds for geometric convergence rates of Markov chains. *The Annals of Applied Probability*, pp. 981–1011, 1994.

Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308 – 2335, 2015. doi: 10.3150/14-BEJ645. URL `https://doi.org/10.3150/14-BEJ645`.

Alexander V Nazin, Boris T Polyak, and Alexandre B Tsybakov. Optimal and robust kernel algorithms for passive stochastic approximation. *IEEE Transactions on Information Theory*, 38(5):1577–1583, 1992.

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

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983. ISBN 9780471103455. URL https://books.google.fr/books?id=6ULvAAAAMAAJ.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. High probability convergence of clipped-SGD under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

Boris Teodorovich Polyak and Yakov Zalmanovich Tsypkin. Adaptive estimation algorithms: convergence, optimality, stability. *Automation and Remote Control*, 40(3):378–389, 1979.

Boris Teodorovich Polyak and Yakov Zalmanovich Tsypkin. Robust pseudogradient adaptation algorithms. *Automation and Remote Control*, 41(10):1404–1409, 1981.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 2018.

Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A robust univariate mean estimator is all you need. In *International Conference on Artificial Intelligence and Statistics*, pp. 4034–4044. PMLR, 2020.

E Price and V VandeLinde. Robust estimation using the robbins-monro stochastic approximation algorithm. *IEEE Transactions on Information Theory*, 25(6):698–704, 1979.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Gareth O Roberts and Richard L Tweedie. Rates of convergence of stochastically monotone and continuous time Markov models. *Journal of Applied Probability*, 37(2):359–373, 2000.

Byron Roe. MiniBooNE particle identification. UCI Machine Learning Repository, 2010. DOI: https://doi.org/10.24432/C5QC87.

Jeffrey S Rosenthal. Convergence rates for Markov chains. *Siam Review*, 37(3):387–405, 1995a.

Jeffrey S Rosenthal. Minorization conditions and convergence rates for markov chain monte carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995b.

Thomas J. Rothenberg, Franklin M. Fisher, and C. B. Tilanus. A note on estimation from a cauchy sample. *Journal of the American Statistical Association*, 59(306):460–463, 1964. doi: 10.1080/01621459.1964.10482170. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1964.10482170.

Peter J Rousseeuw and Mia Hubert. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79, 2011.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pp. 29563–29648. PMLR, 2023.

Prem Seetharaman, Gordon Wichern, Bryan Pardo, and Jonathan Le Roux. Autoclip: Adaptive gradient clipping for source separation networks. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2020.

Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 807–814, 2007.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Conference on Learning Theory*, volume 2, pp. 5, 2009.

Srdjan S Stanković and Branko D Kovačević. Analysis of robust stochastic approximation algorithms for process identification. *Automatica*, 22(4):483–488, 1986.

Che-Ping Tsai, Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. Heavy-tailed streaming statistical estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1251–1282. PMLR, 2022.

John Wilder Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pp. 448–485, 1960.

Andrew V Uzilov, Joshua M Keegan, and David H Mathews. Detection of non-coding rnas on the basis of predicted secondary structure formation free energy change. *BMC bioinformatics*, 7(1):1–30, 2006.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pp. 65–102. PMLR, 2022.

Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. *Advances in Neural Information Processing Systems*, 34:4234–4248, 2021.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 116, 2004.