# Bayesian Robustness: A Nonasymptotic Viewpoint

Kush Bhatia† kushbhatia@berkeley.edu
Yi-An Ma† yianma@berkeley.edu
Anca D. Dragan† anca@berkeley.edu
Peter L. Bartlett†,‡ peter@berkeley.edu
Michael I. Jordan†,‡ jordan@cs.berkeley.edu

†Department of Electrical Engineering and Computer Sciences
‡Department of Statistics, University of California, Berkeley

July 30, 2019

**Abstract**

We study the problem of robustly estimating the posterior distribution for the setting where observed data can be contaminated with potentially adversarial outliers. We propose Rob-ULA, a robust variant of the Unadjusted Langevin Algorithm (ULA), and provide a finite-sample analysis of its sampling distribution. In particular, we show that after $T = \tilde{\mathcal{O}}(d/\varepsilon_{\mathsf{acc}})$ iterations, we can sample from $p_T$ such that $\mathrm{dist}(p_T, p^*) \leq \varepsilon_{\mathsf{acc}} + \tilde{\mathcal{O}}(\epsilon)$, where $\epsilon$ is the fraction of corruptions. We corroborate our theoretical analysis with experiments on both synthetic and real-world data sets for mean estimation, regression and binary classification.

## 1 Introduction

Robustness has been of ongoing interest in both the Bayesian [DF61, BMP+94] and frequentist setting [Tuk60, Hub64, Hub73a] since being introduced by George Box in 1953 [Box53]. The goal is to capture the sensitivity of inferential procedures to the presence of outliers in the data and misspecifications in the modelling assumptions, and to mitigate overly large sensitivity. The Bayesian approach has been focused on capturing possible anomalies in the observed data via the model and in choosing priors that have minimal effect on inferences. The frequentist approach, on the other hand, has focused on the development of estimators that identify and guard against outliers in the data. We refer the reader to [Hub11, Chap 15] for a comprehensive discussion.

The focus on *model robustness* in Bayesian statistics is implemented via sensitivity studies to understand effects of misspecification of the prior distribution [BMP+94, MSLD17] and its propagation towards the posterior [Hub73b]. There is, however, little in the way of a comprehensive formal finite-sample framework for Bayesian robustness. Huber asked "Why there is no finite sample Bayesian robustness theory?" and Kong suggested that such a theory is infeasible in full generality, arguing that it is computationally infeasible to carry out the necessary calculations even in finite spaces.

We address this issue by providing a formal framework for studying Bayesian robustness and by proposing a robust inferential procedure with finite-sample guarantees. We address issues of computational infeasibility by refraining from modelling outlier data explicitly. Instead, we posit that the collected data contains a small fraction of observations which are not explained by the modelling assumptions. This corruption model, termed an $\epsilon$-contamination model, was first studied by Huber in 1964 [Hub64] and has been the subject of recent computationally-focused work in the frequentist setting [DKK+16, LRV16, PSBR18].

Given data corrupted according to an $\epsilon$-contamination model, our goal is to sample from the *clean* posterior distribution $p^*$: the posterior distribution conditioning only on the uncorrupted ("clean") data. Our key idea is to leverage a robust approach for estimating the mean in the context of gradient-based sampling techniques. Our overall procedure is a robust variant of the Unadjusted Langevin Algorithm (ULA) that we refer to as "Rob-ULA." The underlying ULA algorithm and its variants have been used for efficient large-scale Bayesian posterior sampling [RT96, WT11] and their convergence analysis has been a recent topic of interest [Dal17, DM17, CB18, DCWY18, MCJ+18]; see Section 2.2 for a detailed overview. Informally, our main result shows that after $T = \tilde{\mathcal{O}}(d/\varepsilon_{\mathsf{acc}})$ iterations of Rob-ULA, the iterate $\theta_T$ has a distribution $p_T$ such that $\mathrm{dist}(p_T, p^*) \leq \varepsilon_{\mathsf{acc}} + \tilde{\mathcal{O}}(\epsilon)$, where $\epsilon$ is the fraction of corrupted points in the data set.

The remainder of the paper is organized as follows: Section 2 contains a discussion of the related literature, Section 3 discusses relevant background as well as the formal problem setup, and Section 4 describes the proposed algorithm Rob-ULA and states our main theorem regarding its convergence. In Section 5 we discuss the fundamental problems of Bayesian mean estimation and linear regression in the robust setting. Section 6 consists of experimental evaluation of Rob-ULA on synthetic as well as real world data sets and we conclude with Section 7.

## 2 Related Work

In this section, we review related work on robust estimation procedures in both the Bayesian and frequentist settings. We also discuss work on using Langevin dynamics to sample from distributions over continuous spaces.

### 2.1 Robust statistical procedures

There are many threads in the literature on robust estimation and outlier detection [Hub73a, Box53, DF61]. In the frequentist parameter estimation setting, the most commonly studied model is Huber's classical $\epsilon$-contamination model. There has also been a recent focus on an adversarial paradigm that is devoted to developing computationally efficient problem-dependent estimators for mean estimation [LRV16, DKK+16], linear regression [KKM18, BJK15, BJKK17, SBRJ19], and general risk minimization problems [PSBR18, DKK+19]. Particular relevant to our setup are [PSBR18] and [DKK+19] which utilize the robust mean estimators of [LRV16, DKK+16] to robustify gradient-based procedures for empirical risk minimization.

The study of robustness in the Bayesian framework has focused primarily on developing models and priors that have minimal impact on inference. An important line of work has focused on the sensitivity of the posterior distribution to and has led to the development of noninformative priors [BMP+94, MSLD17, MD18]. These methods are orthogonal to those considered in the current paper, as they do not aim to robustify inferential procedures against corruptions in the observed data set. In principle a complete Bayesian model would have the capacity for explaining the outliers present in the data, but this would require performing an integral over all models with a proper prior. Such an approach would generally be computationally intractable.

An important class of procedures for handling outliers in the data set focuses on *reweighing* the data points via a transformation of the likelihood function. Huber [Hub11] considers assigning binary weights to data points and identifies model-dependent procedures to identify outliers. In contrast, Wang et al. [WKB17] consider these weights as latent variables and infers these weight variables along with the parameters of the model. These methods are susceptible to the choice of priors over these weighting variables. An alternate set of robust procedures are based on the idea of *localization* [DF61, WB18]:

each data point is allowed to depend on its own version of the latent variable and these individual latent variables are tied together with the use of a hyperparameter, often fitted using empirical Bayes estimation. Although these methods build on intuitive notions of outlier removal, there is little theoretical understanding of the kind of outliers that these methods can tolerate.

### 2.2 Sampling methods

Turning to sampling methods for posterior inference, there have been various zeroth-order [LS93, MT96] and first-order methods [Erm75, RR98, Nea11] proposed for sampling from distributions over continuous spaces. Our focus in this paper is on the overdamped Langevin MCMC method which was first proposed by Ermak [Erm75] in the context of molecular dynamics applications. Its nonasymptotic convergence (in total variation distance) was first studied by Dalalyan [Dal17] for log-smooth and log-strongly concave distributions. Cheng and Bartlett [CB18] extended the analysis to obtain a similar convergence result in Kullback-Leibler divergence. Such nonasymptotic convergence guarantees are essential to understanding the robustness of computational procedures as they simultaneously capture the dependence of the sampling error on the number of iterations $T$, the dimensionality $d$, and the contamination level $\epsilon$.

## 3 Background

In the section we briefly review relevant background on Bayesian computation and we formally describe our problem setup.

### 3.1 Bayesian modelling

Given parameters $\theta \in \mathbb{R}^d$ and a data set $\mathcal{D} = \{z_1, z_2, \ldots, z_n\}$, we assume that the statistician has specified a prior distribution, $p_\theta(\theta|\alpha)$, and a likelihood, $p(z|\theta)$. We can then form the posterior distribution, $p(\theta|\mathcal{D}, \alpha)$, as follows:

$$p(\theta|\mathcal{D}, \alpha) \propto p_\theta(\theta|\alpha) \cdot \prod_{i=1}^{n} p(z_i|\theta) \; .$$

We are generally concerned with the estimation of some test function $h(\theta)$ under the posterior distribution, which is accomplished in the Bayesian setting by computing a posterior mean:

$$q(h|\mathcal{D}, \alpha) := \int_{\mathbb{R}^d} h(\theta) p(\theta|\mathcal{D}, \alpha) \mathrm{d}\theta \; .$$

In practice, one way of computing this posterior mean is to use a Monte Carlo algorithm to generate a sequence of samples $\{\theta_t\}_{t=1}^T$ and form an estimate $\hat{q}(h|\mathcal{D}, \alpha)$:

$$\hat{q}(h|\mathcal{D}, \alpha) = \frac{1}{T} \sum_{t=1}^{T} h(\theta_t) \; .$$

### 3.2 Unadjusted Langevin Algorithm

We consider a specific Monte Carlo algorithm, the Unadjusted Langevin Algorithm (ULA), for sampling from probability distributions over continuous spaces. Generically, we consider distributions over $\mathbb{R}^d$ of the form

$$p^*(\theta) \propto \exp(-f(\theta)),$$

for a class of functions $f$ which are square integrable. The ULA algorithm starts from some initial point $\theta_0 \in \mathbb{R}^d$ and defines a sequence of parameters, $\{\theta_{k,\eta}\}_{k=1}^T$, according to the following update equation:

$$\theta_{k+1,\eta} = \theta_{k,\eta} - \eta_{k+1}\nabla f(\theta_{k,\eta}) + \sqrt{2\eta_{k+1}}\xi_{k+1}, \tag{1}$$

where $\eta = \{\eta_i > 0\}$ denotes a sequence of step sizes and $\{\xi_i\}_{i\in\mathbb{N}} \sim \mathcal{N}(0, \mathrm{I}_{d\times d})$ are i.i.d. Gaussian vectors. The Markov chain in Equation (1) is the Euler discretization of a continuous-time diffusion process $\{\theta_t\}_{t\geq 0}$ known as the Langevin diffusion. The stochastic differential equation governing the Langevin diffusion is given by

$$\mathrm{d}\theta_t = -\nabla f(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\ , \quad t \geq 0, \tag{2}$$

where $\{B_t\}_{t\geq 0}$ represents a $d$-dimensional Brownian motion. Denoting the distribution of $\theta_{k,\eta}$ by $p_{k,\eta}$, Cheng and Bartlett [CB18] showed that $\mathrm{KL}(p_{k,\eta} \parallel p^*) \leq \epsilon$ after $t = \tilde{\mathcal{O}}(\frac{d}{\epsilon})$ steps for functions $f$ which are smooth and strongly-convex. Specializing to the Bayesian modelling setup we rewrite the posterior distribution as:

$$p(\theta|\mathcal{D}, \alpha) \propto \exp\left(\log(p_\theta(\theta|\alpha)) - \sum_{i=1}^n g_i(\theta)\right)\ ,$$

where $g_i(\theta) := -\log(p(z_i|\theta))$ is the negative log-likelihood corresponding to the $i^{th}$ data point. The ULA algorithm can then be used to form an approximation to the posterior as follows:

$$\theta_{k+1,\eta} = \theta_{k,\eta} - \eta_{k+1}\left(-\nabla\log(p_\theta(\theta_{k,\eta}|\alpha) + \sum_{i=1}^n \nabla g_i(\theta_{k,\eta})\right) + \sqrt{2\eta_{k+1}}\xi_{t+1},$$

where $\eta$ and $\xi_{t+1}$ are the step-size sequence and independent Gaussian noise respectively.

### 3.3 Problem Setup

We turn to a formal treatment of the robustness problem in the Bayesian setting. We consider the $\epsilon$-contamination model introduced by Huber [Hub64] and let the collection of $n$ data points be obtained from the following mixture distribution:

$$z_i \sim (1-\epsilon)P + \epsilon Q\,, \tag{3}$$

where $P$ denotes the true underlying generative distribution while $Q$ is any arbitrary distribution. A data set $\mathcal{D}$ drawn from such a mixture distribution has each data point $z_i$ corrupted adversarially with probability $\epsilon$. We denote by $\mathcal{D}_c$ the subset of data points in $\mathcal{D}$ sampled from the true distribution $P$ and similarly let $\mathcal{D}_a$ denote the subset of data sampled from $Q$. Given data points $\mathcal{D} = \mathcal{D}_c \cup \mathcal{D}_a$, the likelihood function $p(z|\theta)$ and the prior $p_\theta(\theta|\alpha)$, we aim to form a *clean* posterior distribution given by:

$$p(\theta|\mathcal{D}_c, \alpha) \propto p_\theta(\theta|\alpha) \cdot \prod_{i\in\mathcal{D}_c}(z_i|\theta)\ .$$

Accordingly, as in Section 3.1, we would like to *robustly* estimate the mean of the test function $h(\theta)$ under the uncorrupted posterior $p(\theta|\mathcal{D}_c, \alpha)$:

$$q(h|\mathcal{D}, \alpha) := \int_{\mathbb{R}^d} h(\theta)p(\theta|\mathcal{D}_c, \alpha)\mathrm{d}\theta,$$

which we approximate via an estimate:

$$\hat{q}(h|\mathcal{D}_c, \alpha) = \frac{1}{T}\sum_{t=1}^T h(\theta_t^c)\ .$$

**Algorithm 1:** Rob-ULA: Robust Unadjusted Langevin Algorithm

**Input:** Data set $\mathcal{D}$, step-size sequence $\eta$, initial covariance scaling $\beta$, timesteps $T$, prior distribution $p_\theta(\theta|\alpha)$, hyperparameters $\alpha$, likelihood function $p(z|\theta)$, corruption level $\epsilon$

Sample $\theta_0 \sim \mathcal{N}(0, \beta I_d)$

**for** $k = 1, \ldots, T$ **do**

Let $g_i(\theta_{k-1,\eta}) := -\log(p(z_i|\theta_{k-1,\eta}))$ for $i = 1 \ldots n$

$\widehat{\nabla} U_\theta = \text{RobustGradientEstimate}(\{\nabla g_i(\theta_{k-1,\eta})\}_{i=1}^n, \epsilon, d)$

$\theta_{k,\eta} = \theta_{k-1,\eta} - \eta_k (n \cdot \widehat{\nabla} U_\theta - \nabla \log(p_\theta(\theta_{k-1,\eta}|\alpha)) + \sqrt{2\eta_k}\xi_k, \quad \text{where } \xi_k \sim \mathcal{N}(0, I_d)$

**Output:** Iterates $\{\theta_k\}$

In the following section we present an algorithm, Rob-ULA, for generating the sequence of samples $\{\theta_t^c\}_{t=1}^T$ and provide theoretical guarantees on its convergence properties. The main idea is to exploit gradient-based iterative sampling methods, and to leverage a robust mean estimation procedure to compute a robust estimate of the gradient at each iteration.

## 4 Rob-ULA: Robust Unadjusted Langevin Algorithm

We turn to our proposed algorithm, Rob-ULA (Algorithm 1), which aims to solve the robust posterior inference problem defined in Section 3.3. Rob-ULA is a simple modification of the ULA algorithm, described in Section 3.2, where in each iteration instead of using the complete set of data points for computing the gradient, we construct a *robust* estimator of the gradient and update the parameter using this estimate. This robust estimator ensures that the outlier data points do not exert too much influence on the gradient and allow Rob-ULA to obtain samples from a distribution close to the clean posterior distribution:

$$p(\theta|\mathcal{D}_c, \alpha) \propto p_\theta(\theta|\alpha) \cdot \prod_{i \in \mathcal{D}_c} (z_i|\theta) \,. \tag{4}$$

Before proceeding to establish the convergence guarantees for Rob-ULA, we present the robust gradient estimation procedure.

### 4.1 Robust Gradient Estimation

Algorithm 2 describes our robust gradient estimation procedure. Based on the robust mean estimator of Lai et al. [LRV16], it takes as input the gradients of the negative log-likelihoods $\nabla g_i(\theta)$ and outputs an estimate of the *robust mean* of the gradient vectors ($\widehat{\nabla} U_\theta$ in Algorithm 1), assuming a fraction $\epsilon$ of them are arbitrarily corrupted. Algorithm 1 then scales this gradient estimate by the number of samples $n$, to obtain a robust estimate of gradients of the likelihood $\sum_{i=1}^n \nabla g_i(\theta)$.

Note that the model described in Section 3.3 assumes that each data point $z$ is sampled i.i.d. from the mixture distribution $(1-\epsilon)P + \epsilon Q$, where $P$ represents the true generative distribution and $Q$ can be any arbitrary distribution. An application of the Hoeffding bound for Bernoulli random variables shows that with probability at least $1-\delta$, the fraction of corrupted points $\epsilon_n$ in the sampled data set $\mathcal{D}$ satisfy

$$\epsilon - \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)} \leq \epsilon_n \leq \epsilon + \underbrace{\sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}}_{e_n} . \tag{5}$$

**Algorithm 2:** RobustGradientEstimate

**Input:** Sample Gradients $S = \{\nabla g_i(\theta)\}_{i=1}^n$, Corruption Level $\epsilon$, Dimension $d$
$\tilde{S}$ = Outlier Truncation$(S, \epsilon, d)$
**if** $d = 1$ **then**
  $\hat{\mu} \leftarrow \text{mean}(\tilde{S})$
**else**
  $\Sigma_{\tilde{S}} \leftarrow$ sample covariance of $\tilde{S}$
  Let $V$ = span of top $d/2$ principal components of $\Sigma_{\tilde{S}}$ and $W = V^c$
  $S_1 \leftarrow P_V(\tilde{S})$ where $P_V$ is projection onto $V$
  $\hat{\mu}_V \leftarrow$ Robust Gradient Estimator $(S_i, \epsilon, d/2)$
  $\hat{\mu}_W \leftarrow \text{mean}(P_W(\tilde{S}))$
  $\hat{\mu} \leftarrow \hat{\mu}_V + \hat{\mu}_W$
**Output:** Estimate of Robust Gradient: $\hat{\mu}$

**Algorithm 3:** Outlier Truncation

**Input:** Sample Gradients $S = \{\nabla g_i(\theta)\}_{i=1}^n$, Corruption Level $\epsilon$, Dimension $d$
**if** $d = 1$ **then**
  $[a, b] \leftarrow$ smallest interval containing $(1-\epsilon)^2$ fraction of points.
  $\tilde{S} \leftarrow S \cap [a, b]$
**else**
  Let $[S]_i$ be samples with only $i^{th}$ coordinate
  **for** $i = 1, \ldots, d$ **do**
    $a[i] =$ Robust Gradient Estimator$([S]_i, \epsilon, 1)$
  Let $B(r, a)$ be ball of smallest radius centred at $a$ containing $(1-\epsilon)^2$ fraction of points.
  $\tilde{S} \leftarrow S \cap B(r, a)$
**Output:** Points after outlier removal : $\tilde{S}$

For the remainder of the paper, we condition on this high probability event and state our results assuming this event holds. Following the proof strategy of [LRV16] and [PSBR18], we derive a bound on the estimation error of the true average log-likelihood gradient,

$$\left\| \widehat{\nabla} U_\theta - \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} \nabla g_i(\theta) \right\|_2 ,$$

*uniformly* for any value of the iterate $\theta$ in the following lemma. We let $\nabla U_\theta := \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} \nabla g_i(\theta)$ denote the true value of this average log-likelihood gradient.

**Lemma 1 (Robust Gradient Estimation)** *Let $P$ denote the uniform distribution over $\mathcal{D}_c$ and let $P_\theta$ denote the corresponding distribution over $\nabla g_i(\theta)$ with mean given by $\nabla U_\theta$, covariance $\Sigma_\theta$ and fourth moment given by $C_4$. There exists a positive constant $C_1 > 0$ for which the robust mean estimator when instantiated with the contamination level $\gamma := \epsilon + e_n$, returns, with probability $1 - \delta$, an estimate $\widehat{\nabla} U_\theta$ such that for all $\theta \in \mathbb{R}^d$, we have that,*

$$\|\widehat{\nabla} U_\theta - \nabla U_\theta\|_2 \leq C_1 C_4^{\frac{1}{4}} \sqrt{\gamma \log(d) \|\Sigma_\theta\|_2} .$$

**Remark.** Note that Proposition 1 of Prasad et al. [PSBR18] presents a high-probability bound similar to ours which is applicable for a *fixed* parameter $\theta$. Such a bound, however,

does not suffice to ensure convergence of Rob-ULA because the additive Gaussian noise at every iterate requires us to obtain a *uniform* high-probability recovery error bound (see Section 4.2 for details). Lemma 1 establishes a uniform bound for the specific distribution $P$ which is uniform over the clean data $\mathcal{D}_c$. This restriction of the distribution $P$ also allows us to avoid sample-splitting at every iteration of Algorithm 1 which was essential for both [LRV16] and [PSBR18].

In addition, Lemma 1 indicates that irrespective of the sample size $n$, one can estimate the mean of the gradient robustly up to error $\tilde{\mathcal{O}}(\sqrt{\epsilon\|\Sigma_\theta\|_2})$. This implies that at each iteration of Rob-ULA, we incur an error of $\tilde{\mathcal{O}}(n\sqrt{\epsilon\|\Sigma_\theta\|_2})$ since we scale the average gradient estimate by $n$ during the update. In Theorem 2 we show how with an appropriate choice of step size $\eta = \mathcal{O}(1/n)$, one can control the propagation of this bias in the convergence analysis for Rob-ULA.

The detailed proof of Lemma 1 can be found in Appendix A.

### 4.2 Convergence Analysis

In this section, we study the convergence of the proposed algorithm Rob-ULA. For ease of notation, we let $f(\theta;\mathcal{D}) = \sum_{i\in\mathcal{D}} g_i(\theta) - \log(p_\theta(\theta|\alpha))$ and similarly denote the clean and corrupted versions of the function $f(\theta;\mathcal{D}_c)$ and $f(\theta;\mathcal{D}_a)$. The objective of the robust Bayesian posterior inference problem is then to obtain samples from the clean posterior distribution given by $p^*(\theta|\mathcal{D},\alpha) \propto \exp\left(-f(\theta;\mathcal{D}_c)\right)$. For clarity of exposition, we drop the dependence of the posterior distribution on the data set $\mathcal{D}$ as well as the hyperparameters $\alpha$ and let $f(\theta) := f(\theta;\mathcal{D}_c)$.

We quantify the convergence of distribution $p(\theta)$ following a stochastic process to the stationary distribution $p^*(\theta)$ through the Kullback-Leibler divergence, $\mathrm{KL}(p(\theta) \parallel p^*(\theta))$:

$$\mathrm{KL}(p(\theta) \parallel p^*(\theta)) = \int p(\theta) \ln\left(\frac{p(\theta)}{p^*(\theta)}\right) \mathrm{d}\theta \,.$$

We define the Wasserstein distance $\mathrm{W}_2^2(p,q)$ between a pair of distributions $(p,q)$ as:

$$\mathrm{W}_2^2(p,q) := \inf_{\zeta\in\Gamma(p,q)} \int \|x-y\|_2^2 \,\mathrm{d}\zeta(x,y)\,,$$

where $\Gamma(p,q)$ denotes the set of joint distributions such that the first set of coordinates has marginal $p$ and the second set has marginal $q$.

We begin by making the following assumptions on the function $f(\theta)$:

**Assumption 1** (Lipschitz smoothness). The function $f(\theta)$ is $L$-Lipschitz smooth and its Hessian exists for all $\theta \in \mathbb{R}^d$. That is,

$$\|\nabla f(\theta) - \nabla f(\nu)\| \le L\,\|\theta-\nu\|\,,\ \forall\theta,\nu\in\mathbb{R}^d \quad \text{and} \quad \nabla^2 f(\theta) \text{ exists for all } \theta\in\mathbb{R}^d\,.$$

**Assumption 2** (Strong convexity). The function $f(\theta)$ is $m$-strongly convex for all $\theta \in \mathbb{R}^d$. That is,

$$mI \preceq \nabla^2 f(\theta), \quad \forall\theta\in\mathbb{R}^d.$$

We further denote the condition number of the function $f$ as $\kappa = L/m$.

The assumptions of Lipschitz smoothness and strong convexity are standard in both the sampling and optimization literatures. In addition to the assumptions, we define the average Lipschitz constant $\bar{L} = L/n$ and the average strong convexity of $f$ as $\bar{m} = m/n$. We now state our main theorem concerning the convergence guarantees for Rob-ULA.

**Theorem 2 (Main Result)** *Let $p^*(\theta) \propto \exp(-f(\theta))$, where $f$ satisfies Assumptions 1 and 2. Further, assume that the gradient estimates $\widehat{\nabla} f(\theta)$ satisfy*

$$\left\| \nabla f(\theta_{k,\eta}) - \widehat{\nabla} f(\theta_{k,\eta}) \right\|^2 \leq n^2 \epsilon C_R \left\| \Sigma_\theta \right\|_2 \log d \quad \text{and} \quad \left\| \Sigma_\theta \right\|_2 \leq C_{\Sigma,1} \left\| \theta - \tilde{\theta} \right\|^2 + C_{\Sigma,2} \ ,$$

*where $\Sigma_\theta$ is the covariance of uniform distribution on $\nabla g_i(\theta)$ induced by the clean data set $\mathcal{D}_c$, $\tilde{\theta}$ satisfies $\nabla F(\tilde{\theta}) = 0$ and $\epsilon$ is the fraction of corrupted points, satisfying $\epsilon \leq \frac{\bar{m}^2}{4 C_R C_{\Sigma,1} \log d}$. Then the iterates of Rob-ULA, when initialized with $\theta_0 \sim \mathcal{N}\left(0, \frac{1}{L} I_d\right)$ (with corresponding density $p_0$) and step size $\eta \leq \frac{1}{nL}$ (and define $h := n\eta \leq \frac{1}{L}$), satisfy:*

$$\begin{aligned}
\mathrm{W}_2^2(p_{k\eta}, p^*) &\leq \frac{2e^{-\bar{m}kh}}{n\bar{m}} \mathrm{KL}(p_0 \parallel p^*) + 8 \left( C_R C_{\Sigma,2} \frac{\bar{L}^4}{\bar{m}^4} \epsilon \log d + \frac{\bar{L}^4}{\bar{m}^3} \frac{d}{n} \right) h^2 + 4 \frac{\bar{L}^2}{\bar{m}^2} \frac{d}{n} h \\
&\quad + \epsilon \left( \frac{4 C_R C_{\Sigma,2}}{\bar{m}^2} \log d + \frac{8 C_R C_{\Sigma,1}}{\bar{m}^2} \frac{d \log d}{n} \right) \\
&\leq \frac{1}{\bar{m}} \log \frac{\bar{L}}{\bar{m}} \frac{d}{n} e^{-\bar{m}kh} + 8 \left( C_R C_{\Sigma,2} \frac{\bar{L}^4}{\bar{m}^4} \epsilon \log d + \frac{\bar{L}^4}{\bar{m}^3} \frac{d}{n} \right) h^2 + 4 \frac{\bar{L}^2}{\bar{m}^2} \frac{d}{n} h \\
&\quad + \epsilon \left( \frac{4 C_R C_{\Sigma,2}}{\bar{m}^2} \log d + \frac{8 C_R C_{\Sigma,1}}{\bar{m}^2} \frac{d \log d}{n} \right) ,
\end{aligned}$$

*where $p_{k\eta}$ represents the distribution of the iterate $\theta_{k,\eta}$.*

**Remarks.** Before proceeding to the proof of this theorem, a few comments are in order. First observe that the error term consists of three different components:

$$\underbrace{\frac{2e^{-n\bar{m}k\eta}}{n\bar{m}} \mathrm{KL}(p_0 \parallel p^*)}_{(I)} + \underbrace{C \left( \frac{\bar{L}^4}{\bar{m}^4} \epsilon \log d + \frac{\bar{L}^4}{\bar{m}^3} \frac{d}{n} \right) h^2 + 4 \frac{\bar{L}^2}{\bar{m}^2} \frac{d}{n} h}_{(II)} + \underbrace{\frac{C}{\bar{m}^2} \epsilon \log d}_{(III)},$$

where a) term (I) comprises an exponentially decaying dependence (with the number of time-steps $t$) on the initial error $\mathrm{KL}(p_0 \parallel p^*(\theta))$, b) term (II) is a discretization error term and c) term (III) captures the dependence on the fraction of corrupted points $\epsilon$ and vanishes as $\epsilon$ goes to zero.

For any given accuracy $\varepsilon_{\mathsf{acc}}$, if the step size and the number of iterations satisfy:

$$\eta = \mathcal{O}\left( \frac{\varepsilon_{\mathsf{acc}}}{n \kappa \bar{L} d} \right) \quad \text{and} \quad T \geq \mathcal{O}\left( \frac{\bar{L}}{\bar{m}} \log \left( \frac{\mathrm{KL}(p_0 \parallel p^*)}{n \bar{m} \varepsilon_{\mathsf{acc}}} \right) \right),$$

then the error in convergence can be bounded as

$$\mathrm{W}_2^2(p_{T,\eta}, p^*) \leq \varepsilon_{\mathsf{acc}} + \tilde{\mathcal{O}}\left( \frac{\epsilon}{\bar{m}^2} \right).$$

As we show in Section 5, for problems such as Bayesian linear regression and Bayesian mean estimation, the average strong convexity parameter $\bar{m}$ scales independently of the sample size $n$. This implies that the resulting error can be bounded by $\varepsilon_{\mathsf{acc}} + \tilde{\mathcal{O}}(\epsilon)$. While the accuracy can be set to arbitrarily small values which would result in a corresponding increase in the number of time steps, there is a bias term depending on the contamination level $\mathcal{O}(\epsilon)$ which cannot be reduced by either increasing the sample size or by increasing the number of iterations. This is consistent with results in the frequentist literature [BJK15, DKK+16, LRV16, PSBR18], which show that such inconsistency is a result of the adversarial corruptions and in general cannot be avoided.

Lemma 13 in Appendix B presents the following bound on the initial error:

$$\mathrm{KL}(p_0 \parallel p^*) = \int p_0(\mathbf{x}) \log \left( \frac{p_0(\mathbf{x})}{p^*(\mathbf{x})} \right) \mathrm{d}\mathbf{x} \leq \frac{d}{2} \log \frac{\bar{L}}{\bar{m}}.$$

## Proof of Theorem 2

We proceed to a proof of our main convergence result, Theorem 2. We begin by considering the process described by Rob-ULA as a discretization of the Langevin dynamics given by Equation 2, with the following gradient estimate:

$$\Theta_{(k+1)\eta} = \Theta_{k\eta} - \eta\widehat{\nabla} f(\Theta_{k\eta}) + \sqrt{2}(B_{(k+1)\eta} - B_{k\eta}), \tag{6}$$

where $\Theta_{k\eta}$ represents the random variable describing the process at the $k^{th}$ iterate using step size $\eta$. This is equivalent to defining the following stochastic differential equation

$$\mathrm{d}\Theta_t = -\widehat{\nabla} f(\Theta_{kh})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \qquad \text{for} \qquad k\eta < t \leq (k+1)\eta\,. \tag{7}$$

We next state a lemma which provides a bound on the variance of the distribution of the $k^{th}$ iterate.

**Lemma 3** *For $\Theta_t$ following Eq. (7), if $\Theta_0 \sim \mathcal{N}\left(0, \frac{1}{n\bar{L}}\mathrm{I}\right)$, $\epsilon \leq \frac{\bar{m}^2}{4C\log d}$, and $h := n\eta \leq \frac{1}{\bar{L}}$, then for all $k \in \mathbb{N}^+$,*

$$\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|_2^2\right] \leq \frac{4}{\bar{m}^2}C'\epsilon\log d + \frac{4d}{n\bar{m}},$$

*where $C$ is a universal constant and $\tilde{\theta}$ is the fixed point satisfying $\nabla f(\tilde{\theta}) = 0$.*

The proof of this lemma is deferred to Appendix B. Treating Lemma 3 as given, we proceed to the proof of Theorem 2.

We consider the dynamics in Equation 7 within the time range $k\eta < t \leq (k+1)\eta$. From the Girsanov theorem [Oks03] we have that $\Theta_t$ admits a density function $p_t$ with respect to the Lebesgue measure. This density function can also be represented as

$$p_t(\theta) = \int p_{k\eta}(\zeta)p(\theta, t|\zeta, k\eta)\mathrm{d}\zeta,$$

where $p(\theta, t|\zeta, k\eta)$ is the weak solution to the following Kolmogorov forward equation:

$$\frac{\partial p(\theta, t|\zeta, k\eta)}{\partial t} = \nabla^T\big(\nabla p(\theta, t|\zeta, k\eta) + \widehat{\nabla} f(\zeta)p(\theta, t|\zeta, k\eta)\big),$$

where $p(\theta, t|\zeta, k\eta)$ and its derivatives are defined via $P_t(f) = \int f(\theta)p(\theta, t|\zeta, k\eta)\mathrm{d}\theta$ as a functional over the space of smooth bounded functions on $\mathbb{R}^d$ (we refer the readers to [SP14] for more details). As shown by Cheng and Bartlett [CB18], the time derivative of the KL divergence along $p_t$ is given by:

$$\frac{d}{dt}\mathrm{KL}(p_t \parallel p^*) = -\mathbb{E}\left[\left\langle \nabla \ln\left(\frac{p_t(\Theta_t)}{p^*(\Theta_t)}\right), \nabla \ln p_t(\Theta_t) + \widehat{\nabla} f(\Theta_{k\eta})\right\rangle\right],$$

where the expectation is taken with respect to the joint distribution of $\Theta_t$ and $\Theta_{k\eta}$. Hence

$$\begin{aligned}\frac{d}{dt}\mathrm{KL}(p_t \parallel p^*) &\overset{\text{(i)}}{=} -\mathbb{E}\left[\left\langle \nabla \ln\left(\frac{p_t(\Theta_t)}{p^*(\Theta_t)}\right), \nabla \ln\left(\frac{p_t(\Theta_t)}{p^*(\Theta_t)}\right) + \left(\widehat{\nabla} f(\Theta_{k\eta}) - \nabla f(\Theta_t)\right)\right\rangle\right] \\ &= -\mathbb{E}\left[\left\|\nabla \ln\left(\frac{p_t(\Theta_t)}{p^*(\Theta_t)}\right)\right\|^2\right] + \mathbb{E}\left[\left\langle \nabla \ln\left(\frac{p_t(\Theta_t)}{p^*(\Theta_t)}\right), \nabla f(\Theta_t) - \widehat{\nabla} f(\Theta_{k\eta})\right\rangle\right],\end{aligned} \tag{8}$$

where (i) follows from the definition of $p^*(\theta) \propto \exp(-f(\theta))$. We first focus on the second term in the above expression which can be bounded as:

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \nabla \ln\left(\frac{p_t(\Theta_t)}{p^*(\Theta_t)}\right), \nabla f(\Theta_t) - \widehat{\nabla} f(\Theta_{k\eta})\right\rangle\right] \\
&\quad= \mathbb{E}\left[\left\langle \nabla \ln\left(\frac{p_t(\Theta_t)}{p^*(\Theta_t)}\right), (\nabla f(\Theta_t) - \nabla f(\Theta_{k\eta})) + \left(\nabla f(\Theta_{k\eta}) - \widehat{\nabla} f(\Theta_{k\eta})\right)\right\rangle\right] \\
&\quad\overset{\text{(i)}}{\leq} \frac{1}{2}\mathbb{E}\left[\left\|\nabla \ln\left(\frac{p_t(\Theta_t)}{p^*(\Theta_t)}\right)\right\|^2\right] + \mathbb{E}\left[\|\nabla f(\Theta_t) - \nabla f(\Theta_{k\eta})\|^2\right] + \mathbb{E}\left[\left\|\nabla f(\Theta_{k\eta}) - \widehat{\nabla} f(\Theta_{k\eta})\right\|^2\right] \\
&\quad\overset{\text{(ii)}}{\leq} \frac{1}{2}\mathbb{E}\left[\left\|\nabla \ln\left(\frac{p_t(\Theta_t)}{p^*(\Theta_t)}\right)\right\|^2\right] + L^2\mathbb{E}\left[\|\Theta_t - \Theta_{k\eta}\|^2\right] + C_R n^2 \epsilon \left(C_{\Sigma,1}\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|^2\right] + C_{\Sigma,2}\right)\log d,
\end{aligned} \tag{9}$$

where (i) follows by an application of Young's inequality $2a^\top b \leq \|a\|_2^2 + \|b\|_2^2$ and (ii) follows from the point-wise assumption that $\left\|\nabla f(\theta_{k,\eta}) - \widehat{\nabla} f(\theta_{k,\eta})\right\|^2 \leq n^2 \epsilon C_R \left\|\Sigma_\theta\right\|_2 \log d$ and that $\left\|\Sigma_\theta\right\|_2 \leq C_{\Sigma,1}\left\|\theta - \tilde{\theta}\right\|^2 + C_{\Sigma,2}$. Let us define new constant $C_{13} := C_R \cdot C_{\Sigma,1}$ and $C_{14} := C_R \cdot C_{\Sigma,2}$.

Next, we proceed to bound the term $\mathbb{E}\left[\|\Theta_t - \Theta_{k\eta}\|^2\right]$ using Lemma 3. Let us define the variable $\tau := t - k\eta \in (0, \eta]$ and bound the term as:

$$\begin{aligned}
\mathbb{E}\left[\|\Theta_t - \Theta_{k\eta}\|^2\right] &\leq \mathbb{E}\left[\left\|-\nabla f(\Theta_{k\eta})\tau + \sqrt{2}(B_{(k+1)\eta} - B_{k\eta})\right\|^2\right] \\
&\leq \mathbb{E}_{\theta\sim p_{k\eta}}\left[\|\nabla f(\theta)\|^2\right]\tau^2 + 2d\tau \\
&\overset{\text{(i)}}{\leq} \mathbb{E}_{\theta\sim p_{k\eta}}\left[\left\|\theta - \tilde{\theta}\right\|^2\right]\bar{L}^2\nu^2 + 2\frac{d}{n}\nu\,,
\end{aligned} \tag{10}$$

where (i) follows from Assumption 1 and we define $\nu = n\tau$ ($h = n\eta$). Plugging in the bounds obtained in Equations (9) and (10) into Equation (8), we get for $k\eta < t \leq (k+1)\eta$:

$$\begin{aligned}
\frac{d}{dt}\mathrm{KL}(p_t \parallel p^*) &\leq -\frac{1}{2}\mathbb{E}\left[\left\|\nabla \ln\left(\frac{p_t(\Theta_t)}{p^*(\Theta_t)}\right)\right\|^2\right] + n^2\bar{L}^4\nu^2\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|^2\right] + 2nd\bar{L}^2\nu \\
&\quad + C_R n^2\epsilon\left(C_{\Sigma,1}\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|^2\right] + C_{\Sigma,2}\right)\log d \\
&= -\frac{1}{2}\mathbb{E}_{\theta\sim p_t}\left[\left\|\nabla \ln\left(\frac{p_t(\theta)}{p^*(\theta)}\right)\right\|^2\right] + n^2\left(\bar{L}^4\nu^2 + C_{13}\epsilon\log d\right)\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|^2\right] \\
&\quad + n^2\epsilon C_{14}\log d + 2nd\bar{L}^2\nu \\
&\overset{\text{(i)}}{\leq} -m\cdot\mathrm{KL}(p_t \parallel p^*) + n^2\left(\bar{L}^4h^2 + C_{13}\epsilon\log d\right)\left(\frac{4}{\bar{m}^2}C_{14}\epsilon\log d + \frac{4d}{n\bar{m}}\right) \\
&\quad + n^2\epsilon C_{14}\log d + 2n\bar{L}^2 dh \\
&\overset{\text{(ii)}}{\leq} -n\bar{m}\mathrm{KL}(p_t \parallel p^*) + E(h, n, d, \bar{L}, \bar{m}),
\end{aligned}$$

where (i) follows from an application of the log-Sobolev inequality with $m$ being the log-Sobolev constant and (ii) follows from the fact that $\epsilon \leq \dfrac{\bar{m}^2}{4C_{13}\log d}$ and the substitution

$$E(h, n, d, \bar{L}, \bar{m}) = \left(4C_{14}n^2\frac{\bar{L}^4}{\bar{m}^2}\epsilon\log d + 4nd\frac{\bar{L}^4}{\bar{m}}\right)h^2 + 2nd\bar{L}^2 h + 2C_{14}n^2\epsilon\log d + 4C_{13}n\epsilon d\log d.$$

Finally, using Gronwall's inequality we have the following one-step progress equation:

$$\mathrm{KL}(p_{(k+1)\eta} \parallel p^*) - \frac{1}{n\bar{m}}E(h,n,d,\bar{L},\bar{m}) \leq e^{-n\bar{m}\eta}\left(\mathrm{KL}(p_{k\eta} \parallel p^*) - \frac{1}{n\bar{m}}E(h,n,d,\bar{L},\bar{m})\right).$$

Repeated application of this progress inequality leads us to

$$\begin{aligned}\mathrm{KL}(p_{k\eta} \parallel p^*) \leq& e^{-n\bar{m}k\eta}\left(\mathrm{KL}(p_0 \parallel p^*) - \frac{1}{n\bar{m}}E(h,n,d,\bar{L},\bar{m})\right) + \frac{1}{n\bar{m}}E(h,n,d,\bar{L},\bar{m}) \\ \leq& e^{-n\bar{m}k\eta}\mathrm{KL}(p_0 \parallel p^*) + \frac{1}{n\bar{m}}E(h,n,d,\bar{L},\bar{m}).\end{aligned}$$

The final result of the theorem can then be obtained by using Talagrand's inequality [OV00] which states that for the probability distributions $p_{k\eta}$ and $p^*$, we have that

$$\mathrm{W}_2^2(p_{k\eta},p^*) \leq \frac{2}{n\bar{m}}\mathrm{KL}(p_{k\eta} \parallel p^*) \leq \frac{2}{n\bar{m}}e^{-n\bar{m}k\eta}\mathrm{KL}(p_0 \parallel p^*) + \frac{2}{n^2\bar{m}^2}E(h,n,d,\bar{L},\bar{m}),$$

which concludes the proof of the theorem. ■

## 5 Consequences for Mean Estimation and Regression

In this section, we study the fundamental problems of Bayesian mean estimation (Section 5.1) and Bayesian linear regression (Section 5.2) under the Huber $\epsilon$-contamination model.

### 5.1 Robust Bayesian mean estimation

We begin with the *robust Bayesian mean estimation* (RBME) problem and instantiate the convergence guarantees for Rob-ULA (Algorithm 1) for this problem. For simplicity, we study the setup in which the likelihood is Gaussian:

$$p(z|\theta;\Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}}\exp\left(-\frac{1}{2}\|z-\theta\|_{\Sigma^{-1}}^2\right),$$

for a mean vector $\theta \in \mathbb{R}^d$ and a fixed positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. We consider the corresponding conjugate prior over $\theta$ given by

$$p(\theta;\theta_0,\Sigma_0) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_0)}}\exp\left(-\frac{1}{2}\|\theta-\theta_0\|_{\Sigma_0^{-1}}^2\right),$$

where $\theta_0$ is the mean and $\Sigma_0 \succ 0$ is the covariance matrix. The set of parameters $(\theta_0,\Sigma_0)$ are the hyperparameters $\alpha$. Given data points $\mathcal{D} = \{z_i\}_{i=1}^n$ sampled from the Huber contamination model, where $Z_i \overset{\text{i.i.d}}{\sim} (1-\epsilon)P + \epsilon Q$, with $Q$ being an arbitrary adversarially chosen distribution, the objective of the RBME problem is to sample from the posterior induced by the uncorrupted data points,

$$p^* := p(\theta|\mathcal{D}_c;\Sigma,\theta_0,\Sigma_0) \propto \exp\left(-\frac{1}{2}\|\theta-\theta_0\|_{\Sigma_0^{-1}}^2\right)\prod_{i\in\mathcal{D}_c}\exp\left(-\frac{1}{2}\|z_i-\theta\|_{\Sigma^{-1}}^2\right), \tag{11}$$

where $\mathcal{D}_c$ represents the subset of points in $\mathcal{D}$ sampled from the distribution $P$. We note that for data $X_i$ sampled from the Huber contamination model, we have from Equation (5) that with probability at least $1-\delta$,

$$n(1-\epsilon-e_n) \leq |\mathcal{D}_c| \leq n(1-\epsilon+e_n) \quad \text{where} \quad e_n := \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}. \tag{12}$$

Let us denote by $\bar{\epsilon} := \epsilon + e_n$ and by $\underline{\epsilon} := \epsilon - e_n$ the corresponding upper and lower bounds on the number of corrupted data points. Following the notation in Section 4, the function $f(\theta)$ for this problem is given by:

$$f(\theta) = \frac{1}{2}\left( \|\theta - \theta_0\|^2_{\Sigma_0^{-1}} + \sum_{i \in \mathcal{D}_c} \|z_i - \theta\|^2_{\Sigma^{-1}} \right). \tag{13}$$

Also, we define the corresponding function $g_i(\theta)$ from Section 3 as well as the sample mean and covariance for the clean data points.

$$g_i(\theta) := \frac{1}{2}\|z_i - \theta\|^2_{\Sigma^{-1}}, \quad \mu_z := \frac{1}{|\mathcal{D}_c|}\sum_{i \in \mathcal{D}_c} z_i, \quad \Sigma_z := \frac{1}{|\mathcal{D}_c|}\sum_{i \in \mathcal{D}_c} (z_i - \mu_z)(z_i - \mu_z)^\top.$$

The following corollary instantiates the guarantees of Theorem 2 for the specific function $f(\theta)$ defined for the RBME problem.

**Corollary 4** *Consider the RBME problem with posterior given by Equation (11) and data sampled from the Huber model. Then the iterates of Rob-ULA with step size $\eta \leq \frac{1}{n\bar{L}}$ and $h := n\eta$ satisfy:*

$$\mathrm{W}_2^2(p_{k\eta}, p^*) \leq \frac{2e^{-n\bar{m}k\eta}}{n\bar{m}}\mathrm{KL}(p_0 \parallel p^*) + C\left( C_{\Sigma,2}\frac{\bar{L}^4}{\bar{m}^4}\epsilon \log d + \frac{\bar{L}^4}{\bar{m}^3}\frac{d}{n} \right) h^2 + 4\frac{\bar{L}^2}{\bar{m}^2}\frac{d}{n}h + C\cdot\epsilon\frac{C_{\Sigma,2}}{\bar{m}^2}\log d\,,$$

*where $C$ is a constant depending on the fourth moment of the clean data $\mathcal{D}_c$ and*

$$\bar{m} = \left( \frac{(1-\bar{\epsilon})}{\lambda_{max}(\Sigma)} + \frac{1}{n\lambda_{max}(\Sigma_0)} \right), \quad \bar{L} = \left( \frac{(1-\underline{\epsilon})}{\lambda_{min}(\Sigma)} + \frac{1}{n\lambda_{min}(\Sigma_0)} \right), \quad \textit{and,} \quad C_{\Sigma,2} = \frac{\lambda_{max}(\Sigma_z)}{\lambda_{min}(\Sigma)}\,,$$

*with probability at least $1 - \delta$ for $\epsilon < 1/2$.*

**Remark.** Observe that the step-size parameter $h$ is independent of $n$ while both $\bar{L}$ and $\bar{m}$ are asymptotically independent of the sample size $n$. The above bound shows that for an appropriately chosen step size one can obtain samples from a distribution which is $\tilde{\mathcal{O}}(\epsilon)$ away from the true distribution $p^*$. The number of iterations required to obtain such a sample scales linearly with the number of samples $n$, the dimension $d$ and the condition number $\kappa = \frac{\bar{L}}{\bar{m}}$.

### 5.2 Robust Bayesian linear regression

We turn to the *robust Bayesian Linear Regression* (RBLR) problem. For this problem, we let the data set $\mathcal{D} = \{z_i = (x_i, y_i)\}_{i=1}^n$ be such that $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ be the covariate vectors and response variables sampled from the Huber contamination model. Note that the Huber contamination model is on the variable $z_i$ and hence allows for corruption in both the features $x_i$ as well as the response variables $y_i$. In addition, we assume that there exists a vector $\theta^*$ such that

$$y_i = \langle x_i, \theta^* \rangle + z_i,$$

where $z_i \sim \mathcal{N}(0, \sigma^2)$ are sampled independently of $x_i$. This assumption is for simplifying the presentation and in general one can work with $\theta^*$ which is the best linear approximation to the data. For the RBLR problem, we consider likelihood functions of the form:

$$p((x,y)|\theta;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left( -\frac{1}{2\sigma^2}(y - \langle x, \theta \rangle)^2 \right),$$

for a fixed variance parameter $\sigma^2$. Also, we consider a Gaussian prior over the parameter $\theta$,

$$p(\theta; \theta_0, \Sigma_0) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_0)}}\exp\left( -\frac{1}{2}\|\theta - \theta_0\|^2_{\Sigma_0^{-1}} \right),$$

for some fixed mean vector $\theta_0$ and positive-definite covariance matrix $\Sigma_0$ which form the set of hyperparameters $\alpha$. Given a data set $\mathcal{D}$ sampled from the Huber $\epsilon$-contamination model, the objective of the RBLR problem is to sample from the posterior induced by the uncorrupted set of data points,

$$p(\theta|\mathcal{D}_c;\sigma,\theta_0,\Sigma_0) \propto \exp\left(-\frac{1}{2}\|\theta-\theta_0\|^2_{\Sigma_0^{-1}}\right)\prod_{i\in\mathcal{D}_c}\exp\left(-\frac{1}{2\sigma^2}(y_i-\langle x_i,\theta\rangle)^2\right). \tag{14}$$

Following a similar calculation to that in Section 5.1, we have that with probability at least $1-\delta$,

$$n(1-\bar{\epsilon}) \le |\mathcal{D}_c| \le n(1-\underline{\epsilon}). \tag{15}$$

The corresponding function $f(\theta)$ for the RBLR problem is then defined to be

$$f(\theta) = \frac{1}{2}\left(\|\theta-\theta_0\|^2_{\Sigma_0^{-1}} + \frac{1}{\sigma^2}\sum_{i\in\mathcal{D}_c}(y_i-\langle x_i,\theta\rangle)^2\right). \tag{16}$$

We denote by $\theta_{\text{reg}}$ the estimator which minimizes the function $U(\theta)$, and is given by:

$$\theta_{\text{reg}} := \left(\Sigma^{-1}+\frac{1}{\sigma^2}X_c^\top X_c\right)^{-1}\left(\frac{1}{\sigma^2}X_c^\top y_c + \Sigma^{-1}\theta_0\right),$$

where $X_c \in \mathbb{R}^{n_c\times d}$ represents the set of covariate vectors of the clean data points and $y_c \in \mathbb{R}^{n_c}$ represents the corresponding response values. In addition, we define the following functions required for the analysis of the robust gradient estimator,

$$g_i(\theta) := \frac{1}{2\sigma^2}(y_i-\langle x_i,\theta\rangle)^2, \quad \mu_x := \frac{1}{|\mathcal{D}_c|}\sum_{i\in\mathcal{D}_c}x_i, \quad \widetilde{\Sigma}_x := \frac{1}{|\mathcal{D}_c|}\sum_{i\in\mathcal{D}_c}x_i x_i^\top.$$

We make the following moment assumptions on the covariates in the clean data set $\mathcal{D}_c$.

**Assumption 3** (Positive-definite data covariance). The unnormalized data covariance matrix is positive definite: $\widetilde{\Sigma}_x \succ 0$.

**Assumption 4** (Bounded fourth moment). The data satisfies a bounded fourth moment condition, i.e., for every unit vector $v$, we have that

$$\mathbb{E}_{x\sim_u\mathcal{D}_c}\left[(v^\top x)^4\right] \le C_{x,4}\left(\mathbb{E}_{x\sim_u\mathcal{D}_c}\left[(v^\top x)^2\right]\right)^2,$$

for come constant $C_{x,4}$.

Note that these assumptions are satisfied with high probability if say, each $x_i \in \mathcal{D}_c$ is sampled i.i.d. from the standard normal distribution. The following corollary then instantiates the guarantees of Theorem 2 for the specific function $f(\theta)$ defined for the RBLR problem in Equation (16).

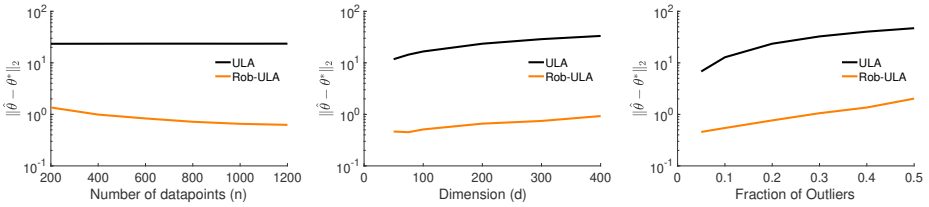
**Corollary 5** *Consider the RBLR problem described above with posterior given by Equation (14) and data sampled from the Huber model. Then the iterates of Rob-ULA with step size $\eta \le \frac{1}{n\bar{L}}$ and $h = n\eta$ satisfy:*

$$\begin{aligned}\mathrm{W}_2^2(p_{k\eta},p^*) \le{}& \frac{2e^{-n\bar{m}k\eta}}{n\bar{m}}\mathrm{KL}(p_0 \parallel p^*) + C\left(C_{\Sigma,2}\frac{\bar{L}^4}{\bar{m}^4}\epsilon\log d + \frac{\bar{L}^4}{\bar{m}^3}\frac{d}{n}\right)h^2 + 4\frac{\bar{L}^2}{\bar{m}^2}\frac{d}{n}h \\ &+ C\epsilon\left(\frac{C_{\Sigma,2}}{\bar{m}^2}\log d + \frac{C_{\Sigma,1}}{\bar{m}^2}\frac{d\log d}{n}\right),\end{aligned}$$

*with probability at least $1-\delta$ for $\epsilon \le \frac{\bar{m}^2}{4CC_{\Sigma,1}\log d}$. The constant $C$ depends on $C_{x,4}$ from Assumption 4 and the remaining parameters are defined as:*

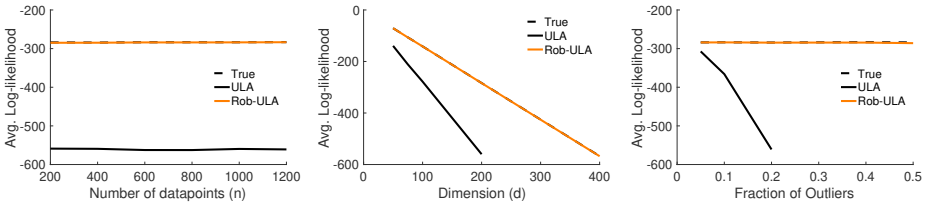
$$\bar{m} = (1-\bar{\epsilon})\lambda_{\min}(\widetilde{\Sigma}_x) + \frac{1}{n\lambda_{\max}(\Sigma_0)}, \quad \bar{L} = (1-\underline{\epsilon})\lambda_{\max}(\widetilde{\Sigma}_x) + \frac{1}{n\lambda_{\min}(\Sigma_0)}, \quad C_{\Sigma,1} = 2\sqrt{8C_{x,4}}\cdot\|\widetilde{\Sigma}_x\|_2$$

$$C_{\Sigma,2} = \sqrt{C_{x,4}}\|\widetilde{\Sigma}_x\|_2 + 2\sqrt{8C_{x,4}}\cdot\|\widetilde{\Sigma}_x\|_2\cdot\|\theta^*-\theta_{\text{reg}}\|_2^2 + \frac{(8C_{z,4})^{\frac{1}{2}}\cdot\sigma^2}{n^{\frac{1}{4}}}\log\left(\frac{e^2}{\delta}\right) + \sqrt{24}\sigma^2.$$




(a) (b) (c)

Figure 1: Robust Bayesian Mean Estimation (Parameter Estimation): Rob-ULA recovers the underlying parameter with smaller error as compared with the vanilla ULA. The recovery error increases with increasing dimension and fraction of outliers.




(a) (b) (c)

Figure 2: Robust Bayesian Mean Estimation (Average log-likelihood): Rob-ULA has average log-likelihood close to the true underlying parameter. The log-likelihood values for ULA become large and negative for higher dimensions and fraction of outliers and do not show up in plots (b) and (c).

**Remark.** As in the RBME problem, the quantities $h$, $\bar{L}$ and $\bar{m}$ are asymptotically independent of the sample size $n$. However, the guarantees above hold only for a value of $\epsilon \leq \tilde{\mathcal{O}}(\frac{1}{\kappa(\Sigma_x)})$, that is, they depend on the condition number of the covariate distribution. Such a dependence seems inherent to the problem of linear regression since the adversary is allowed to corrupt the covariate vector arbitrarily.

## 6 Experiments

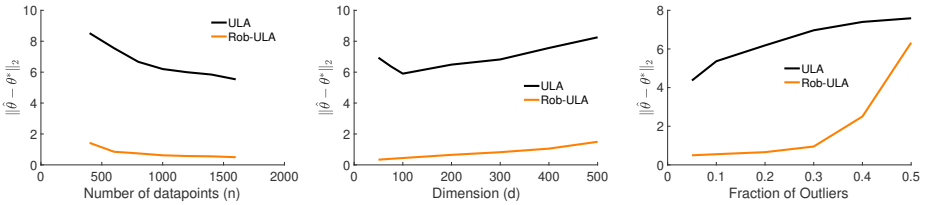
In this section, we compare the performance of the proposed Rob-ULA with the non-robust variant ULA. We first compare them on synthetic data sets for the problem of Bayesian mean estimation and Bayesian linear regression in order to understand the variation in performance as a function of the problem parameters. In Section 6.2, we perform experiments comparing the algorithms on some real-world binary classification data sets using logistic regression.

### 6.1 Synthetic data sets

#### 6.1.1 Robust Bayesian Mean Estimation

In this section, we focus on experiments related to the robust Bayesian mean estimation problem described in Section 5.1. We begin by describing the experimental setup before proceeding to a discussion of the experimental findings.

**Experiment setup.** The mean vector $\theta \in \mathbb{R}^d$ was sampled as a uniform distribution independently in each coordinate over the interval $[0, 1]$. The clean samples $z_i$ were then obtained independently from $\mathcal{N}(\theta, I)$. The corrupted distribution $Q$ was chosen as the Gaussian distribution $\mathcal{N}(\theta_c, I)$ with mean given by $\theta_c = \theta + \theta_{\text{cor}}$ where each entry of $\theta_{\text{cor}}$




(a) (b) (c)

Figure 3: Robust Bayesian Linear Regression (Parameter Estimation): Rob-ULA recovers the underlying parameter with smaller error as compared with the vanilla ULA. The recovery error increases with increasing dimension and fraction of outliers. For $\epsilon \approx 0.5$, the performance of Rob-ULA and ULA become quite similar in terms of recovery guarantees.

is sampled i.i.d. from the uniform distribution over $[0, 10]$. The default parameters were set as follows: number of clean samples $n_c = 1000$, dimension $d = 200$ and fraction of corruption $\epsilon = 0.2$. In every experiment, one of the parameters was varied keeping the others fixed. For both Rob-ULA and ULA, the burn-in period was set to 300 samples and a total of $n_{\mathsf{samp}} = 1000$ samples were collected following the burn-in period. Each experiment was repeated for 10 runs and we report the mean performance of the methods across these runs.

**Recovery guarantees.** Figures 1 and 2 compare the performance of the algorithms for the mean estimation problem. Figure 1 shows the variation in parameter recovery error, $\|\hat{\theta} - \theta\|_2$ where $\hat{\theta} = \frac{1}{n_{\mathsf{samp}}} \sum_i \theta_i$ is the average of the collected samples. Figure 1(a) shows that with increasing number of data points, the error in Rob-ULA's estimate decreases until it starts to saturate. In addition, with an increasing dimension and fraction of outliers, the error in estimation for both Rob-ULA and ULA increases. This is consistent with Theorem 2. Figure 2 studies the variation in average log-likelihood of the average estimate $\hat{\theta}$ on a held-out test set. We also plot the likelihood values obtained by plugging in the true parameter (dotted line). The samples output by Rob-ULA have likelihood values identical to the true underlying parameter while those of ULA are much lower for all experiments. Note that ULA fails to have finite likelihood values (up to Matlab precision) for dimensions $d > 200$ and fraction of corruption $\epsilon > 0.2$ and hence have been omitted from the curves.

### 6.1.2 Robust Bayesian Linear Regression

In this section, we discuss the robust Bayesian linear regression problem described in Section 5.2.

**Experiment setup.** The true parameter $\theta^*$ was selected similarly to the mean vector in the RBME experiments. The $x$ were sampled i.i.d. from $\mathcal{N}(0, I)$ and the corresponding response variable were set as $y = x^\top \theta^* + \mathcal{N}(0, 1)$. For the corrupted distribution, each coordinate of the feature vector $x$ was sampled i.i.d. from a $\chi^2$ distribution and the corresponding response variables were set as $y = x^\top \theta^* + \mathrm{Unif}[0, 10]$. The default parameters were set as follows: number of clean samples $n_c = 1000$, dimension $d = 200$ and fraction of corruption $\epsilon = 0.2$. In every experiment, one of the parameters was varied keeping the others fixed. For both Rob-ULA and ULA, the burn-in period was set to 100 samples and a total of $n_{\mathsf{samp}} = 300$ samples were collected. Each experiment was repeated for 10 runs and we report the mean performance of the methods across these runs.

**Performance guarantees.** Figure 3 shows the performance of Rob-ULA and ULA on the linear regression problem in terms of parameter recovery. Note that similar to the mean estimation setup, the error curves for Rob-ULA are lower than those for ULA as we vary number of data points, the dimensionality of the problem and the fraction of corruptions. The error shows a decreasing trend with increasing number of samples but increases with increasing dimension and fraction of outliers, showing that the robust problems indeed become harder in higher-dimensional spaces and with a larger fraction of outliers.

### 6.2 Real-world data sets

In this section, we explore the performance of Rob-ULA on several real-world binary classification data sets obtained from the UCI repository [DG17]. We use a logistic regression model. While technically this model does not fall within the scope of Theorem 2 (primarily because of the strong-convexity assumption), we find that the experimental results are nonetheless consistent with the theory. We begin by providing details on the data sets used and then proceed to the experimental observations. For all the experiments in the section, the standard normal distribution was chosen as the prior.

**Data sets.** The logistic regression experiments were carried out with the following publicly available binary classification data sets: a) Astro [HCL+03], b) Phishing [MTM12], c) Breast-Cancer [MT96], d) Diabetes [DG17] and e) German Credit [DG17]. We normalized the features to scale between $[-1, 1]$ and used 70% of the available data for training purposes and the remaining 30% for testing purposes. For cases where we were required to tune hyperparameters, we used 20% of the train data for validation purposes. Once the hyperparameters were fixed, we retrained the model with the complete training set. In order to understand the effects of corruptions in these data sets, we *manually* add corruptions to the training subset in the form of label flips (for randomly chosen data points) for the Breast-Cancer, Diabetes and German Credit data set. The experiments with Astro and Phishing data sets are with the original *uncorrupted* data sets.

**Evaluation Metric.** For the above data sets, since the true underlying logistic parameter is unknown, we evaluate the algorithms using the log-likelihoods on the test set. For all data sets, we show two plots: plot (a) displays log-likelihood per data point in the test set, sorted in descending order for both algorithms and plot (b) shows a histogram of log-likelihoods. Plot (a) helps understand trends in prediction likelihoods by showing how the prediction quality degrades while Plot (b) provides an understanding of how the likelihoods concentrate.

#### 6.2.1 Binary Classification

Figure 4 compares the performance of Rob-ULA on a binary classification task with no corruptions in the training set for the Astro and Phishing data sets. In both figures, Rob-ULA is seen to perform better than vanilla ULA: the histogram of likelihoods is more concentrated towards the origin. These data sets are not linearly separable and hence the logistic model may provide a poor fit to the data; Rob-ULA exploits this fact and focuses on a subset of points which it can fit well. This allows it to perform better for a larger range of points as compared to vanilla ULA. These experiments show that if there is model misspecification and the chosen model doesn't fit the complete data, the robust model might fit data selectively and give better confidence bounds for those data points.

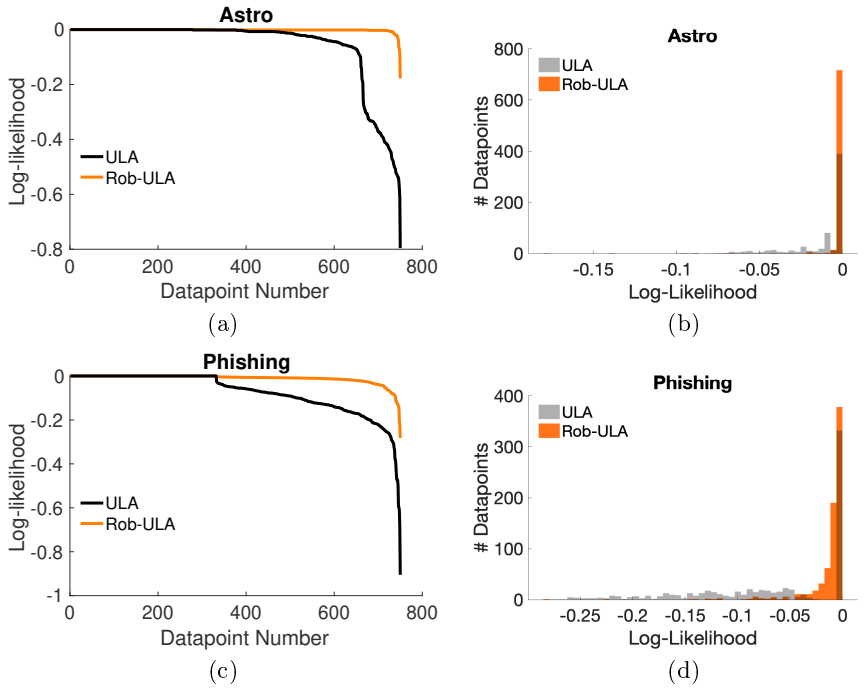



Figure 4: Astro and Phishing Data set (Uncorrupted): Rob-ULA finds solutions which have better test log-likelihood performance across data points as compared to vanilla ULA. Surprisingly, even while ignoring a certain fraction of the data set, the performance of the Rob-ULA does not degrade in those regions of space.

### 6.2.2 Binary Classification with Label Flips

Figure 5 shows the performance of Rob-ULA for the Breast-Cancer ($\epsilon = 0.10$), German Credit ($\epsilon = 0.10$) and Diabetes ($\epsilon = 0.15$) data sets respectively. For these data sets, we manually added corruptions via label flips. For all three data sets, we see a similar trend in likelihood plots: Rob-ULA performs better than vanilla ULA for a majority of data points but its prediction quality decreases for the tail points. This can also be seen in the three histogram plots (part (b) of the respective figures) wherein Rob-ULA has likelihoods extending to larger negative values. This particular behavior can be attributed to the fact that Rob-ULA is unable to learn effective representations in the space where label flips were added since it chooses to ignore those data points. Hence, in the region corresponding to the uncorrupted points, Rob-ULA achieves higher likelihood values than vanilla ULA but for the corrupted regions, the performance of Rob-ULA degrades slightly. This behavior was consistently seen for varying levels of $\epsilon$ with minor shifts in the likelihood curves for different corruption levels.

## 7 Conclusions

We have discussed the problem of robustness to adversarial outliers in a Bayesian framework and proposed Rob-ULA, a robust extension of the classical Unadjusted Langevin algorithm. We obtain nonasymptotic convergence guarantees for Rob-ULA.

We identify multiple directions for future work. On the statistical side, it would be interesting to extend the robustness guarantees of Rob-ULA for statistical models which do not fall within the scope of our current assumptions, notably the case of nonconvex

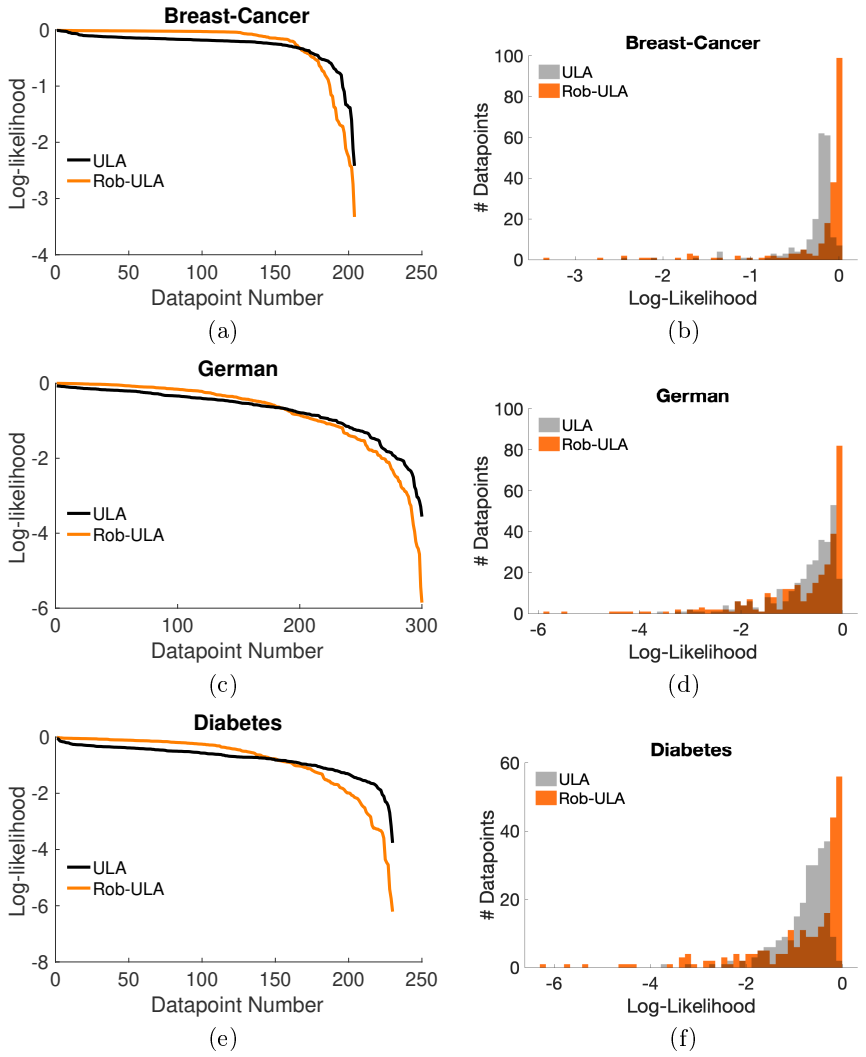



Figure 5: Binary classification with label flips: Breast-Cancer ($\epsilon = 0.10$), German Credit ($\epsilon = 0.10$) and Diabetes ($\epsilon = 0.15$). The plots show that while Rob-ULA is able to achieve high log-likelihood values for a vast majority of the data points, there is a small fraction of the points on which ULA performs better. This behavior can be attributed to Rob-ULA ignoring certain data points during its run and not generalizing well within the subspace spanned by them.

likelihood functions. On the computational side, an important question to understand is whether one can accelerate the convergence of Rob-ULA in the presence of outliers.

## Acknowledgments

This work was done in part while KB, YM and PLB were visiting the Simons Institute for the Theory of Computing.

## References


[BJK15] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

[BJKK17] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2110–2119, 2017.

[BMP+94] James O Berger, Elías Moreno, Luis Raul Pericchi, et al. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

[Box53] George EP Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

[CB18] Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83, pages 186–211. PMLR, 2018.

[Dal17] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[DCWY18] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! *arXiv preprint arXiv:1801.02309*, 2018.

[DF61] Bruno De Finetti. The Bayesian approach to the rejection of outliers. In *Proceedings of the Fourth Berkeley Symposium on Probability and Statistics*, volume 1, pages 199–210. University of California Press Berkeley, 1961.

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science*, pages 655–664. IEEE, 2016.

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1596–1606. PMLR, 2019.

[DM17] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Annals of Applied Probability*, 27(3):1551–1587, 06 2017.

[Erm75] Donald L Ermak. A computer simulation of charged particles in solution. I. Technique and equilibrium properties. *Journal of Chemical Physics*, 62(10):4189–4196, 1975.

[HCL+03] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[Hub64] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[Hub73a] Peter J Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

[Hub73b] Peter J Huber. The use of Choquet capacities in statistics. *Bulletin of the. International. Statistical. Institute*, 45(4):181–191, 1973.

[Hub11] Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[KKM18] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. *arXiv preprint arXiv:1803.03241*, 2018.

[LRV16] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *IEEE 57th Foundations of Computer Science*, pages 665–674. IEEE, 2016.

[LS93] László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures and Algorithms*, 4(4):359–412, 1993.

[MCJ+18] Yi-An Ma, Yuansi Chen, Chi Jin, Nicholas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. arXiv:1811.08413, 2018.

[MD18] Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, pages 1–31, 2018.

[MSLD17] Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B Dunson. Robust and scalable Bayes via a median of subset posterior measures. *The Journal of Machine Learning Research*, 18(1):4488–4527, 2017.

[MT96] Kerrie L Mengersen and Richard L Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 1996.

[MTM12] Rami M Mohammad, Fadi Thabtah, and Lee McCluskey. An assessment of features related to phishing websites using an automated technique. In *International Conference for Internet Technology and Secured Transactions*, pages 492–497. IEEE, 2012.

[Nea11] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

[Oks03] Bernd Oksendal. *Stochastic Differential Equations*. Springer, 6 edition, 2003.

[OV00] Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[PSBR18] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[RR98] Gareth Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[RT96] Gareth Roberts and Richard Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[SBRJ19] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. *arXiv preprint arXiv:1903.08192*, 2019.

[SP14] René L Schilling and Lothar Partzsch. *Brownian Motion: An Introduction to Stochastic Processes*. De Gruyter, 2nd edition, 2014.

[SS12] Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 437–446. SIAM, 2012.

[Tuk60] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

[WB18] Chong Wang and David M Blei. A general method for robust Bayesian modeling. *Bayesian Analysis*, 13:1163–1191, 2018.

[WKB17] Yixin Wang, Alp Kucukelbir, and David M Blei. Robust probabilistic modeling with Bayesian data reweighting. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3646–3655. PMLR, 2017.

[WT11] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688. Omnipress, 2011.

# A Proof of Lemma 1

We begin by obtaining a bound on the performance of Algorithm 2 in the one-dimensional setting and use this as a building block towards the proof for the general $d$-dimensional setting. For ease of exposition, we denote by $\mu_\theta$ and $\hat{\mu}$ the mean gradient $\nabla U_\theta$ and its estimate $\widehat{\nabla} U_\theta$ respectively.

## A.1 Proof for 1 dimensional setting

We begin by analyzing Algorithm 2 for the case when $d = 1$.

**Lemma 6** *Let $P_\theta$ denote the empirical distribution on $\mathcal{D}_c$ in $\mathbb{R}$ with mean $\mu_\theta$, variance $\sigma$ and with fourth moment constant $C_4$. Let $\eta$ be the fraction of corruption in the samples in $\mathcal{D}$. Then Algorithm 2 returns an estimate of the mean $\mu$ such that, for a universal constant $C$,*

$$|\hat{\mu} - \mu_\theta| \leq C C_4^{\frac{1}{4}} \sigma \eta^{\frac{3}{4}}.$$

**Proof** Let $I_{1-\eta}$ be the interval around the true sample mean $\mu_\theta$ containing $1-\eta$ fraction of $\mathcal{D}_c$. Using the bounded fourth moment assumption on $P_\theta$, we have that,

$$\text{length}(I_{1-\eta}) \leq \frac{C_4^{\frac{1}{4}} \sigma}{\eta^{\frac{1}{4}}}. \tag{17}$$

Let $\tilde{S}$ be the set of smallest interval containing $(1-\eta)^2$ fraction of all the points and let this interval be denoted by $\tilde{I}$. Now, we have that $\text{length}(\tilde{I}) \leq \text{length}(I_{1-\eta})$ since the chosen interval has the smallest length. Further, $\tilde{S}$ contains at least $1-3\eta$ fraction of $\mathcal{D}_c$.

For any value of $\eta < 1/6$, we have that the intervals $\tilde{I}$ and $I$ must overlap. Therefore, adversarially corrupted points in the set $\tilde{S}$ are within distance $2 \cdot \text{length}(I_{1-\eta})$ of the true parameter $\mu_\theta$. We now bound the deviation of the estimate $\hat{\mu}$ from $\mu_\theta$ by controlling the sources of error:

**a. Error due to points in $\tilde{S}$ from $\mathcal{D}_a$:** Since there can be at most $\eta$ fraction of corrupted points in our selected sample and each of them within distance $2 \cdot \text{length}(I_{1-\eta})$, their total contribution to the deviation is upper bounded by $2\eta \cdot \text{length}(I_{1-\eta})$.

**b. Error due to points in $\tilde{S}$ from $\mathcal{D}_c$:** Define $\mathcal{A}$ to be the event that a point of $\mathcal{D}_c$ is present in the set $\tilde{S}$. From our discussion above, we have that $P(A) > 1 - 3\eta > 1/2$. Using Lemma 3.11 [LRV16], we have that there exits a constant $C$ such that

$$|\mathbb{E}[X|A] - \mathbb{E}[X]| \leq C C_4^{\frac{1}{4}} \sigma \eta^{\frac{3}{4}}.$$

Combining the analysis of parts (a) and (b) above, we get that,

$$|\hat{\mu} - \mu_\theta| \leq 2\eta \cdot \text{length}(I_{1-\eta}) + C C_4^{\frac{1}{4}} \sigma \eta^{\frac{3}{4}}.$$

Plugging in the bound for $\text{length}(I_{1-\eta})$ completes the proof. ■

## A.2 Proof for d-dimensional setting

We now proceed to prove the robustness properties of Algorithm 2 for the general $d$-dimensional setting. Through the course of this section, we let $\tilde{S}$ be the set of points returned after the outlier truncation procedure with $\tilde{S}_c$ being the clean points and $\tilde{S}_a$ being the adversarially corrupted points. Also, we denote by $\mu_{\tilde{S}} := \text{mean}(\tilde{S})$, $\mu_{\tilde{S}_c} := \text{mean}(\tilde{S}_c)$ and $\mu_{\tilde{S}_a} := \text{mean}(\tilde{S}_a)$ the corresponding mean vectors of the relevant subsets.

**Lemma 7** *Let $P_\theta$ be the empirical distribution over $\mathcal{D}_c$ in $\mathbb{R}^d$ with mean $\mu_\theta$, covariance matrix $\Sigma_\theta$ and with fourth moment constant $C_4$. Let $\eta$ be the fraction of corrupted data points in $\mathcal{D}$. Then, we can obtain a vector $a \in \mathbb{R}^d$ such that for a constant $C$,*

$$\|a - \mu_\theta\|_2 \leq C C_4^{\frac{1}{4}} \sqrt{tr(\Sigma_\theta)} \eta^{\frac{3}{4}}.$$

**Proof** Let $e_1, \ldots, e_d$ be the $d$ canonical basis vectors. Projecting the problem onto these vectors and solving in each direction independently using the method for one dimension, we obtain the bound above by using Lemma 6 separately in each dimension. ■

**Lemma 8** *Let $\eta$ denote the fraction of outliers in $\mathcal{D}$. After the outlier truncation procedure, for every point in the returned set $\tilde{S}$, we have that for a constant $C$,*

$$\|x - \mu_\theta\|_2 \leq C C_4^{\frac{1}{4}} \left( \frac{\sqrt{d\|\Sigma_\theta\|_2}}{\eta^{\frac{1}{4}}} + \sqrt{tr(\Sigma_\theta)} \eta^{\frac{3}{4}} \right).$$

**Proof** Let $B^* = \mathcal{B}(\mu_\theta, r_1^*)$ for $r_1^* = \frac{C C_4^{\frac{1}{4}}}{\eta^{\frac{1}{4}}} \sqrt{d\|\Sigma_\theta\|_2}$ be the $\ell_2$ ball of radius $r_1^*$ around $\mu_\theta$. In order to bound the fraction of clean points in $B^*$, observe that,

$$P(\|x - \mu_\theta\|_2^2 \geq (r_1^*)^2) \leq \frac{\eta \mathbb{E}\left[(\right] \|x - \mu_\theta\|_2^4)}{C_4 d^2 \|\Sigma_\theta\|_2^2}, \tag{18}$$

where the probability is with respect to the empirical distribution over $\mathcal{D}$. Now, $\mathbb{E}\left[(\right] \|x - \mu_\theta\|_2^4 \leq d^2 \max_i \mathbb{E}\left[(\right] (x - \mu_\theta)_i^4) \leq C_4 d^2 \|\Sigma_\theta\|_2^2$. Plugging this value in Equation 18, we get that at least $1 - \eta$ fraction of the points of $\mathcal{D}_c$ lie in $B^*$.

Using Lemma 7, we have that for $r_2^* = C C_4^{\frac{1}{4}} \sqrt{\text{tr}(\Sigma_\theta)} \eta^{\frac{3}{4}}$, there are at least $(1 - \eta)$ fraction of good points at a distance $r_1^* + r_2^*$ away from $a$. For $\eta < 1/6$, we have that the chosen ball of points around $\tilde{S}$ and $B^*$ must overlap. Therefore, minimum radius ball has radius at most $r_1^* + r_2^*$ and when combined with the bound from Lemma 7 and triangle inequality, we get that,

$$\|x - \mu_\theta\|_2 \leq C C_4^{\frac{1}{4}} \left( \frac{\sqrt{d\|\Sigma_\theta\|_2}}{\eta^{\frac{1}{4}}} + \sqrt{\text{tr}(\Sigma_\theta)} \eta^{\frac{3}{4}} \right),$$

which completes the proof. ■

**Lemma 9** *Let $\eta$ be the fraction of corrupted points in $\mathcal{D}$. Then we have that after the outlier truncation step of Algorithm 2, for a constant $C$,*

$$\|\mu_{\tilde{S}_c} - \mu_\theta\|_2 \leq C C_4^{\frac{1}{4}} \eta^{\frac{3}{4}} \sqrt{\|\Sigma_\theta\|_2} \quad \text{and} \quad \|\Sigma_{\tilde{S}_c}\|_2 \leq \|\Sigma_{\tilde{S}_c} - \Sigma_\theta\|_2 + \|\Sigma_\theta\|_2 \overset{\zeta_1}{\leq} (C\eta + 1)\|\Sigma_\theta\|_2.$$

**Proof** We first consider the bound on the mean shift and then proceed with the bound on the covariance matrix.

**Mean Shift Bound:** Let $\mathcal{A}$ be the event that a point $x \in \mathcal{D}_c$ is not removed by the outlier truncation procedure. Then using Lemma 3.11 [[LRV16]] for $\eta < 1/6$ for the random variable $X = x^\top \frac{\mu_{\tilde{S}_c} - \mu_\theta}{\|\mu_{\tilde{S}_c} - \mu_\theta\|_2}$ for $x \sim \mathcal{D}_c$, we have that,

$$\|\mu_{\tilde{S}_c} - \mu_\theta\|_2 \leq CC_4^{\frac{1}{4}} \eta^{\frac{3}{4}} \sqrt{\|\Sigma_\theta\|_2}.$$

**Covariance Matrix Bound:** Consider the following decomposition for bounding the spectral norm of $\Sigma_{\tilde{S}_c}$:

$$\|\Sigma_{\tilde{S}_c}\|_2 \leq \|\Sigma_{\tilde{S}_c} - \Sigma_\theta\|_2 + \|\Sigma_\theta\|_2 \overset{\zeta_1}{\leq} (C\eta + 1)\|\Sigma_\theta\|_2,$$

where $\zeta_1$ follows by using Corollary 3.13 [LRV16] for the same event $\mathcal{A}$ as above in the mean shift bound. ■

**Lemma 10** *$P_W$ is the projection operator on the bottom $d/2$ eigenvectors of the matrix $\Sigma_{\tilde{S}}$. Then, for a constant $C$, we have that,*

$$\|\eta P_W \delta_\mu\|_2^2 \leq \eta((C\eta + 1) + CC_4^{\frac{1}{2}} \eta^{\frac{1}{2}})\|\Sigma_\theta\|_2,$$

*where $\delta_\mu := \mu_{\tilde{S}_a} - \mu_{\tilde{S}_c}$.*

**Proof** Consider the matrix $\Sigma_{\tilde{S}}$. It can be decomposed as:

$$\Sigma_{\tilde{S}} = \underbrace{(1-\eta)\Sigma_{\tilde{S}_c}}_{\Sigma_1} + \underbrace{\eta\Sigma_{\tilde{S}_a} + \eta(1-\eta)\delta_\mu\delta_\mu^\top}_{\Sigma_2}.$$

By Weyl's inequality, we have that,

$$\lambda_{d/2}(\Sigma_{\tilde{S}}) \leq \lambda_1(\Sigma_1) + \lambda_{d/2}(\Sigma_2).$$

We begin by first controlling the term $\lambda_{d/2}(\Sigma_2)$. We have that,

$$\lambda_{d/2}(\Sigma_2) \leq \frac{\mathrm{tr}(\Sigma_2)}{d/2} \overset{\zeta_1}{\leq} C\eta \frac{(r_1^*)^2 + (r_2^*)^2}{d/2} \leq CC_4^{\frac{1}{2}} \|\Sigma_\theta\|_2 \eta^{\frac{1}{2}}, \tag{19}$$

where $\zeta_1$ follows by using the fact that all selected points are in a ball of radius $r_1^* + r_2^*$ where $r_1^*$ and $r_2^*$ are as defined in Lemma 8. Next we consider the term $\lambda_1(\Sigma_1)$ as follows,

$$\lambda_1(\Sigma_1) \overset{\zeta_1}{\leq} (1-\eta)(C\eta + 1)\|\Sigma_\theta\|_2, \tag{20}$$

where $\zeta_1$ follows from using the bound in Lemma 9. Combining Equations (19) and (20), we have that,

$$\lambda_{d/2}(\Sigma_{\tilde{S}}) \leq (1-\eta)(C\eta + 1)\|\Sigma_\theta\|_2 + CC_4^{\frac{1}{2}} \|\Sigma_\theta\|_2 \eta^{\frac{1}{2}}.$$

Using the fact that $P_W$ is the projection operator on the bottom $d/2$ eigenvectors of the matrix $\Sigma_{\tilde{S}}$, we have that,

$$P_W^\top \Sigma_{\tilde{S}} P_W \preceq ((1-\eta)(C\eta + 1) + CC_4^{\frac{1}{2}} \eta^{\frac{1}{2}})\|\Sigma_\theta\|_2 I.$$

Following some algebraic manipulation as in [LRV16], we obtain that

$$\|\eta P_W \delta_\mu\|_2^2 \leq \eta((C\eta + 1) + CC_4^{\frac{1}{2}} \eta^{\frac{1}{2}})\|\Sigma_\theta\|_2.$$

which completes the proof. ■

## Proof of Lemma 1

Let $\tilde{S}$ be the subset of samples returned by the outlier truncation procedure and let $\tilde{S}_c$ be the set of clean points contained in $\tilde{S}$. Then, we have,

$$\begin{aligned}\|\hat{\mu}-\mu_\theta\|_2^2 &\overset{\zeta_1}{=} \|P_W(\hat{\mu}-\mu_\theta)\|_2^2 + \|P_V(\hat{\mu}-\mu_\theta)\|_2^2 \\ &\overset{\zeta_2}{\leq} 2\|P_W(\hat{\mu}-\hat{\mu}_{\tilde{S}_c})\|_2^2 + 2\|P_W(\hat{\mu}_{\tilde{S}_c}-\mu_\theta)\|_2^2 + \|\hat{\mu}_V - P_V\mu_\theta)\|_2^2 \\ &\overset{\zeta_3}{\leq} 2\|P_W(\hat{\mu}-\hat{\mu}_{\tilde{S}_c})\|_2^2 + 2\|(\hat{\mu}_{\tilde{S}_c}-\mu_\theta)\|_2^2 + \underbrace{\|\hat{\mu}_V - P_V\mu_\theta)\|_2^2}_{(I)}, \end{aligned} \tag{21}$$

where $\zeta_1$ follows from the orthogonality of the spaces $V$ and $W$, $\zeta_2$ follows from using triangle inequality and $\zeta_3$ follows from contraction of projection operators. Note that $(I)$ is a problem defined on the subspace $V$ which is of ambient dimension $d/2$ and is solved recursively by Algorithm 2. Thus, one can recursively bound the overall error of the algorithm as,

$$\|\hat{\mu}-\mu_\theta\|_2^2 \leq \left(2\|P_W(\hat{\mu}-\hat{\mu}_{\tilde{S}_c})\|_2^2 + 2\|(\hat{\mu}_{\tilde{S}_c}-\mu_\theta)\|_2^2\right)(1+\log d). \tag{22}$$

Using Lemma 9 and Lemma 10, we can bound the above error as,

$$\|\hat{\mu}-\mu_\theta\|_2 \leq CC_4^{\frac{1}{4}}\sqrt{\eta\log(d)\|\Sigma_\theta\|_2},$$

which completes the proof of the lemma. ■

## B Convergence of Rob-ULA: Proofs for Auxiliary Lemmas

**Lemma 11** *For $\Theta_t$ following Eq. (7), if the initial iterate $\Theta_0 \sim \mathcal{N}\left(0, \frac{1}{n\bar{L}}\mathrm{I}\right)$, the fraction of corruption $\epsilon \leq \frac{\bar{m}^2}{4C_RC_{\Sigma,1}\log d}$, and scaled step-size $h = n\eta \leq \frac{1}{\bar{L}}$, then for all $k \in \mathbb{N}^+$,*

$$\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}^*\right\|_2^2\right] \leq \frac{4C_RC_{\Sigma,2}}{\bar{m}^2}\epsilon\log d + \frac{4d}{n\bar{m}}.$$

**Proof** Consider first the initial iterate for $k=0$. The distribution $p_0$ satisfy $\mathbb{E}_{\theta\sim p_0}\left[\|\theta\|_2^2\right] = \frac{d}{n\bar{L}} \leq \frac{4}{\bar{m}^2}C_{14}\epsilon\log d + \frac{4d}{n\bar{m}}$. We will prove the lemma statement by strong induction.

In the induction hypothesis step, assume that for some $k \geq 0$, for all $t = 0, \eta, \ldots, k\eta$, $\mathbb{E}_{\theta\sim p_t}\left[\|\theta\|_2^2\right] \leq \frac{4}{\bar{m}^2}C_RC_{\Sigma,2}\epsilon\log d + \frac{4d}{n\bar{m}}$. We consider obtaining a bound on $\mathbb{E}_{\theta\sim p_{(k+1)\eta}}\left[\|\theta\|_2^2\right]$, where $p_t$ follows Equation 7, for $t \in (k\eta, (k+1)\eta]$ (denote $\tau = t - k\eta \in (0, h]$):

$$\Theta_t = \Theta_{k\eta} - \widehat{\nabla} f(\Theta_{k\eta})\tau + \sqrt{2}(B_t - B_{k\eta}). \tag{23}$$

Given the above equation, we consider the bound on $\mathbb{E}\left[\left\|\Theta_t - \tilde{\theta}\right\|_2^2\right]$ for some $t \in (k\eta, (k+1)\eta]$ as follows:

$$\begin{aligned}\mathbb{E}\left[\left\|\Theta_t - \tilde{\theta}\right\|_2^2\right] &= \mathbb{E}\left[\left\|\left(\Theta_{k\eta} - \tilde{\theta}\right) - \widehat{\nabla} f(\Theta_{k\eta})\tau + \sqrt{2}(B_t - B_{k\eta})\right\|_2^2\right] \\ &= \mathbb{E}\left[\left\|\left(\Theta_{k\eta} - \tilde{\theta}\right) - \widehat{\nabla} f(\Theta_{k\eta})\tau\right\|_2^2\right] + 2d\tau.\end{aligned}$$

We next define $\nu = n\tau$ and obtain a bound on the term $\mathbb{E}\left[\left\|\left(\Theta_{k\eta} - \tilde{\theta}\right) - \widehat{\nabla} f(\Theta_{k\eta})\tau\right\|_2^2\right]$:

$$\begin{aligned}
\mathbb{E}\left[\left\|\left(\Theta_{k\eta} - \tilde{\theta}\right) - \widehat{\nabla} f(\Theta_{k\eta})\tau\right\|_2^2\right] &= \mathbb{E}\left[\left\|\left(\Theta_{k\eta} - \tilde{\theta}\right) - \frac{1}{n}\widehat{\nabla} f(\Theta_{k\eta})\nu\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|\left(\Theta_{k\eta} - \tilde{\theta}\right) - \frac{1}{n}\nabla f(\Theta_{k\eta})\nu + \frac{1}{n}\nabla f(\Theta_{k\eta})\nu - \frac{1}{n}\widehat{\nabla} f(\Theta_{k\eta})\nu\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|\left(\Theta_{k\eta} - \tilde{\theta}\right) - \frac{1}{n}\nabla f(\Theta_{k\eta})\nu\right\|_2^2\right] + \frac{\nu^2}{n^2}\mathbb{E}\left[\left\|\nabla f(\Theta_{k\eta}) - \widehat{\nabla} f(\Theta_{k\eta})\right\|_2^2\right] \\
&\quad + 2\nu\mathbb{E}\left[\left\langle\left(\Theta_{k\eta} - \tilde{\theta}\right) - \frac{1}{n}\nabla f(\Theta_{k\eta})\nu, \frac{1}{n}\nabla f(\Theta_{k\eta}) - \frac{1}{n}\widehat{\nabla} f(\Theta_{k\eta})\right\rangle\right] \\
&\overset{\text{(i)}}{\leq} (1 + \bar{m}\nu)\,\mathbb{E}\left[\left\|\left(\Theta_{k\eta} - \tilde{\theta}\right) - \frac{1}{n}\nabla f(\Theta_{k\eta})\nu\right\|_2^2\right] \\
&\quad + \left(\frac{\nu}{\bar{m}} + \nu^2\right)\frac{1}{n^2}\mathbb{E}\left[\left\|\nabla f(\Theta_{k\eta}) - \widehat{\nabla} f(\Theta_{k\eta})\right\|_2^2\right],
\end{aligned}$$

where (i) follows by an application of Cauchy–Schwarz inequality. Next, using the assumption on the robust estimation of the gradient from Theorem 2, we have that

$$\frac{1}{n^2}\mathbb{E}\left[\left\|\nabla f(\Theta_{k\eta}) - \widehat{\nabla} f(\Theta_{k\eta})\right\|_2^2\right] \leq C_R C_{\Sigma,1}\epsilon\log d \cdot \mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|^2\right] + C_R C_{\Sigma,2}\epsilon\log d,$$

and further simplifying the above using Lemma 12, we obtain

$$\mathbb{E}\left[\left\|(\Theta_{k\eta} - \mathbf{x}^*) - \frac{1}{n}\nabla U(\Theta_{k\eta})\nu\right\|_2^2\right] \leq (1 - \bar{m}\nu)^2\mathbb{E}\left[\left\|\Theta_{k\eta} - \mathbf{x}^*\right\|^2\right]. \tag{24}$$

Therefore, since $\nu \leq \frac{1}{\bar{L}}$ and the corruption factor$\epsilon \leq \frac{\bar{m}^2}{4C_R C_{\Sigma,1}\log d}$, we have that

$$\begin{aligned}
\mathbb{E}\left[\left\|\left(\Theta_{k\eta} - \tilde{\theta}\right) - \frac{1}{n}\widehat{\nabla} f(\Theta_{k\eta})\nu\right\|_2^2\right] &\leq \left(1 - \bar{m}^2\nu^2\right)(1 - \bar{m}\nu)\,\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|^2\right] \\
&\quad + \left(\frac{\nu}{\bar{m}} + \nu^2\right)\left(C_{13}\epsilon\log d\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|^2\right] + C_{14}\epsilon\log d\right) \\
&\leq \left(1 - \bar{m}\nu + \frac{2\nu}{\bar{m}}C_{13}\epsilon\log d\right)\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|^2\right] + \frac{2\nu}{\bar{m}}C_{14}\epsilon\log d \\
&\leq \left(1 - \frac{\bar{m}\nu}{2}\right)\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|^2\right] + \frac{2\nu}{\bar{m}}C_{14}\epsilon\log d,
\end{aligned}$$

where we have defined the constants $C_{13} := C_R C_{\Sigma,1}$ and $C_{14} := C_R C_{\Sigma,2}$. Using the above bounds, we have that

$$\mathbb{E}\left[\left\|\Theta_t - \tilde{\theta}\right\|_2^2\right] \leq \left(1 - \frac{\bar{m}\nu}{2}\right)\mathbb{E}\left[\left\|\Theta_{k\eta} - \mathbf{x}^*\right\|^2\right] + \frac{2\nu}{\bar{m}}C_{14}\epsilon\log d + \frac{2\nu}{n}d.$$

Note that $\nu \leq \frac{1}{\bar{L}} \leq \frac{1}{\bar{m}}$ and $\mathbb{E}\left[\left\|\Theta_{k\eta} - \tilde{\theta}\right\|^2\right] \leq \frac{4C_{14}}{\bar{m}^2}\epsilon\log d + \frac{4d}{n\bar{m}} = \frac{2}{\bar{m}}\left(\frac{2}{\bar{m}}C_{14}\epsilon\log d + \frac{2}{n}d\right)$.

Combining these, we can obtain the final bound stated in the lemma as follows:

$$\begin{aligned}\mathbb{E}\left[\left\|\Theta_t-\tilde{\theta}\right\|_2^2\right] &\leq \left(1-\frac{\bar{m}\nu}{2}\right)\frac{2}{\bar{m}}\left(\frac{2}{\bar{m}}C_{14}\epsilon\log d+\frac{2}{n}d\right)+\frac{2\nu}{\bar{m}}C_{14}\epsilon\log d+\frac{2\nu}{n}d\\ &\leq \frac{2}{\bar{m}}\left(\frac{2}{\bar{m}}C_{14}\epsilon\log d+\frac{2}{n}d\right)\\ &= \left(\frac{4}{\bar{m}^2}C_{14}\epsilon\log d+\frac{4d}{n\bar{m}}\right),\end{aligned}$$

for any $t\in(k\eta,(k+1)\eta]$. This concludes the proof. ■

We now prove the Lemma 12 which was used in the proof of Lemma 11.

**Lemma 12** *For $\tau\leq\frac{1}{\bar{L}}$, and $\frac{1}{n}\nabla f(\theta)$ being $\bar{m}$ strongly convex and $\bar{L}$ Lipschitz smooth, we have*

$$\left\|(\theta-\tilde{\theta})-\frac{1}{n}\nabla f(\theta)\nu\right\|_2^2\leq(1-\bar{m}\nu)^2\left\|\theta-\tilde{\theta}\right\|_2^2.$$

**Proof** To bound $\left\|\theta-\frac{1}{n}\nabla f(\theta)\nu\right\|_2^2$, we consider the following function: $F(\theta)=\frac{1}{2}\|\theta\|_2^2-\frac{1}{n}f(\theta)\nu$. First note strong convexity and Lipschitz smoothness of $\frac{1}{n}\nabla f(\theta)$ implies that $\bar{m}\mathrm{I}\preceq\frac{1}{n}\nabla^2 f(\theta)\preceq\bar{L}\mathrm{I}$. Thus with $\nu\leq\frac{1}{\bar{L}}$, we have that

$$(1-\nu\bar{L})\mathrm{I}\preceq\nabla^2F(\theta)\preceq(1-\nu\bar{m})\mathrm{I}\quad\forall\,\theta\in\mathbb{R}^d.$$

Note that the point $\tilde{\theta}$ satisfies $\nabla f(\tilde{\theta})=0$. Using this we have that:

$$\begin{aligned}\left\|(\theta-\tilde{\theta})-\frac{1}{n}\nabla f(\theta)\nu\right\|_2^2 &= \left\|\left(\theta-\frac{1}{n}\nabla f(\theta)\nu\right)-\left(\tilde{\theta}-\frac{1}{n}\nabla f(\tilde{\theta})\nu\right)\right\|_2^2\\ &= \left\|\int_0^1\nabla^2F\left(\lambda\theta+(1-\lambda)\tilde{\theta}\right)\mathrm{d}\lambda(\theta-\tilde{\theta})\right\|_2^2\\ &\overset{\text{(i)}}{\leq}(1-\bar{m}\nu)^2\left\|\theta-\tilde{\theta}\right\|_2^2,\end{aligned}$$

where (i) follows from the Lipschitz-smoothness of $F$. ■

**Lemma 13 (Bound on Initial Error)** *If we let the initial iterate $\Theta_0$ have distribution given by*

$$p_0(\theta)=\left(\frac{L}{2\pi}\right)^{d/2}\exp\left(-\frac{L}{2}||\theta||^2\right)$$

*and $p^*$ following Assumptions 1–2, then the initial error $\mathrm{KL}(p_0\parallel p^*)$ is bounded as:*

$$\mathrm{KL}(p_0\parallel p^*)=\int p_0(\theta)\ln\left(\frac{p_0(\theta)}{p^*(\theta)}\right)\mathrm{d}\theta\leq\frac{d}{2}\ln\frac{L}{m}.$$

**Proof** We want to bound $\mathrm{KL}(p_0 \parallel p^*) = \int p_0(\theta) \ln\left(\frac{p_0(\theta)}{p^*(\theta)}\right) \mathrm{d}\theta$, where $p^*(\theta) \propto e^{-f(\theta)}$. First define $\bar{f}(\theta) = f(\theta) - f(\theta^*)$, where $\theta^*$ is the minimum of $f$. Then

$$p^*(\theta) = \frac{\exp\left(-\bar{f}(\theta)\right)}{\int \exp\left(-\bar{f}(\theta)\right) \mathrm{d}\theta}.$$

By Assumptions 1 and 2, we have that $\frac{m}{2}\|\theta\|^2 \leq \bar{f}(\theta) \leq \frac{L}{2}\|\theta\|^2$, $\forall \theta \in \mathbb{R}^d$. Therefore,

$$\begin{aligned} -\ln p^*(\theta) &= \bar{f}(\theta) + \ln \int \exp\left(-\bar{f}(\theta)\right) \mathrm{d}\theta \\ &\overset{\text{(i)}}{\leq} \frac{L}{2}\|\theta\|^2 + \ln \int \exp\left(-\frac{m}{2}\|\theta\|^2\right) \mathrm{d}\theta \\ &= \frac{L}{2}\|\theta\|^2 + \frac{d}{2}\ln\frac{2\pi}{m}, \end{aligned}$$

where (i) follows from using the Lipschitz-smoothness of $\bar{f}$. Hence

$$-\int p_0(\theta) \ln p^*(\theta) \mathrm{d}\theta \leq \frac{d}{2}\ln\frac{2\pi}{m} + \frac{d}{2}.$$

We can also calculate similarly that

$$\int p_0(\theta) \ln p_0(\theta) \mathrm{d}\theta = -\frac{d}{2}\ln\frac{2\pi}{L} - \frac{d}{2}.$$

Combining the above, we get that

$$\mathrm{KL}(p_0 \parallel p^*) = \int p_0(\theta) \ln p_0(\theta) \mathrm{d}\theta - \int p_0(\theta) \ln p^*(\theta) \mathrm{d}\theta \leq \frac{d}{2}\ln\frac{L}{m} = \frac{d}{2}\ln\frac{\bar{L}}{\bar{m}}.$$

This concludes the proof of the statement. ■

# C Proofs for Mean Estimation and Regression

## C.1 Proof of Corollary 4

Throughout this proof, we condition on the high probability event described by Equation (5). We proceed to obtain a bound on the strong-convexity parameter $m$ and Lipschitz smoothness parameter $L$ for the function $f(\theta)$ defined above in Equation (13). The gradient $\nabla f(\theta)$ is given by

$$\nabla f(\theta) = \Sigma_0^{-1}(\theta - \theta_0) + \sum_{i \in \mathcal{D}_c} \Sigma^{-1}(\theta - z_i),$$

and the corresponding hessian $\nabla^2 f(\theta)$ is given by,

$$\nabla^2 f(\theta) = \Sigma_0^{-1} + |\mathcal{D}_c| \cdot \Sigma^{-1}.$$

Since both the matrices $\Sigma_0$ and $\Sigma$ are positive-definite, we have the following bounds for the parameters $m$ and $L$:

$$\underbrace{\left(\frac{n(1-\bar{\epsilon})}{\lambda_{\max}(\Sigma)} + \frac{1}{\lambda_{\max}(\Sigma_0)}\right)}_{m} I \preceq \nabla^2 f(\theta) \preceq \underbrace{\left(\frac{n(1-\underline{\epsilon})}{\lambda_{\min}(\Sigma)} + \frac{1}{\lambda_{\min}(\Sigma_0)}\right)}_{L} I. \tag{25}$$

We now obtain a bound on the covariance $\Sigma_\theta$ of the gradients $\nabla g_i(\theta)$ using the empirical distribution of the points in $\mathcal{D}_c$ as follows:

$$\begin{aligned}\Sigma_\theta &= \frac{1}{|\mathcal{D}_c|}\sum_{i\in\mathcal{D}_c}\left(\nabla g_i(\theta) - \mathbb{E}_{z\sim_u\mathcal{D}_c}\nabla g_z(\theta)\right)\left(\nabla g_i(\theta) - \mathbb{E}_{z\sim_u\mathcal{D}_c}\nabla g_z(\theta)\right)^\top \\ &= \Sigma^{-1}\cdot\frac{1}{|D_c|}\sum_{i\in\mathcal{D}_c}(z_i-\mu_z)(z_i-\mu_z)^\top \\ &= \Sigma^{-1}\Sigma_z, \end{aligned} \tag{26}$$

where $z\sim_u\mathcal{D}_c$ denotes data sampled uniformly from the data set $\mathcal{D}_c$. Thus, from the above equation, we get that $C_{\Sigma,1}=0$ and $C_{\Sigma,2}=\frac{\lambda_{\max}(\Sigma_z)}{\lambda_{\min}(\Sigma)}$. Plugging in these values in Theorem 2 gives us the desired bound. ■

## D Proof of Corollary 5

We condition on the high probability event described by Equation (15). We begin by obtaining a bound on the strong-convexity parameter $m$ and Lipschitz smoothness parameter $L$ for the function $f(\theta)$ defined above in Equation (16). The gradient $\nabla f(\theta)$ is given by

$$\nabla_\theta f(\theta) = \Sigma_0^{-1}(\theta-\theta_0) + \frac{1}{\sigma^2}\sum_{i\in\mathcal{D}_c}\left(x_i\langle x_i,\theta\rangle - y_i x_i\right),$$

and the corresponding hessian $\nabla^2 f(\theta)$ is given by,

$$\nabla^2 f(\theta) = \Sigma_0^{-1} + \frac{|\mathcal{D}_c|}{\sigma^2}\cdot\widetilde{\Sigma}_x.$$

Since both the matrices $\Sigma_0$ and $\widetilde{\Sigma}_x$ are positive-definite, we have the following bounds for the parameters $m$ and $L$:

$$\underbrace{\left(n(1-\bar{\epsilon})\lambda_{\min}(\widetilde{\Sigma}_x)+\frac{1}{\lambda_{\max}(\Sigma_0)}\right)}_{m} I \preceq \nabla^2_\theta f(\theta) \preceq \underbrace{\left(n(1-\underline{\epsilon})\lambda_{\max}(\widetilde{\Sigma}_x)+\frac{1}{\lambda_{\min}(\Sigma_0)}\right)}_{L} I. \tag{27}$$

We now proceed to obtain the bound on the spectral norm of the covariance matrix $\Sigma_\theta$ of the gradients $\nabla g_i(\theta) = x_i(\langle x_i,\theta\rangle - y_i)$ using the empirical distribution of the points in $\mathcal{D}_c$. We use the notation $\mathbb{E}_{\mathcal{D}_c}$ to denote $\mathbb{E}_{(x,y)\sim_u\mathcal{D}_c}$, the sampling of pairs $(\mathbf{x},y)$ uniformly from the clean data set $\mathcal{D}_c$.

$$\begin{aligned}\|\Sigma_\theta\|_2 &= \sup_{v\in\mathbb{S}^{d-1}} v^\top\left(\mathbb{E}_{\mathcal{D}_c}\left[\nabla g_i(\theta)\nabla g_i(\theta)^\top\right] - \mathbb{E}_{\mathcal{D}_c}\left[\nabla g_i(\theta)\right]\mathbb{E}_{\mathcal{D}_c}\left[\nabla g_i(\theta)\right]^\top\right)v \\ &\overset{\text{(i)}}{\leq} \sup_{v\in\mathbb{S}^{d-1}} v^\top\left(\mathbb{E}_{\mathcal{D}_c}\left[\nabla g_i(\theta)\nabla g_i(\theta)^\top\right]\right)v \\ &= \sup_{v\in\mathbb{S}^{d-1}}\mathbb{E}_{\mathcal{D}_c}\left[(v^\top x)^2(\langle x,\theta\rangle - y)^2\right] \\ &\overset{\text{(ii)}}{\leq} \sup_{v\in\mathbb{S}^{d-1}}\sqrt{\mathbb{E}_{\mathcal{D}_c}\left[(v^\top x)^4\right]}\sqrt{\mathbb{E}_{\mathcal{D}_c}\left[(\langle x,\theta\rangle - y)^4\right]}, \end{aligned} \tag{28}$$

where (i) follows from the fact that $\mathbb{E}_{\mathcal{D}_c}\left[\nabla g_i(\theta)\right]\mathbb{E}_{\mathcal{D}_c}\left[\nabla g_i(\theta)\right]^\top \succeq 0$ and (ii) follows from the Cauchy-Schwarz inequality. We now obtain a bound on the two expectations in the above equation.

**Bound on** $\mathbb{E}_{\mathcal{D}_c}\left[(v^\top x)^4\right]$: This term can be bounded using bounded fourth moment assumption (Assumption 4) as follows:

$$\mathbb{E}_{\mathcal{D}_c}\left[(v^\top x)^4\right] \leq C_{x,4}\left(\mathbb{E}_{x\sim_u \mathcal{D}_c}\left[(v^\top x)^2\right]\right)^2 \leq C_{x,4}\|\widetilde{\Sigma}_x\|_2^2. \tag{29}$$

**Bound on** $\mathbb{E}_{\mathcal{D}_c}\left[(\langle x,\theta\rangle - y)^4\right]$: We simplify this term by using the modelling assumption on the data $(x,y)$ and then proceed to bound this using the $c_r$-inequality.

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_c}\left[(\langle x,\theta\rangle - y)^4\right] &= \mathbb{E}_{\mathcal{D}_c}\left[(\langle x,\Delta_\theta\rangle - z)^4\right] \\ &\overset{\text{(i)}}{\leq} 8\cdot\left(\mathbb{E}_{\mathcal{D}_c}\left[(\langle x,\Delta_\theta\rangle)^4\right] + \mathbb{E}_{\mathcal{D}_c}\left[z^4\right]\right) \\ &\overset{\text{(ii)}}{\leq} 8\cdot\left(C_{x,4}\|\widetilde{\Sigma}_x\|_2^2\|\Delta_\theta\|_2^4 + \mathbb{E}_{\mathcal{D}_c}\left[z^4\right]\right),\end{aligned}$$

where $\Delta_\theta := \theta - \theta^*$, (i) follows from using the $c_r$-inequality $\mathbb{E}|X+Y|^r \leq 2^{r-1}\left(\mathbb{E}|X|^r + \mathbb{E}|Y|^r\right)$ and (ii) follows from Assumption 2 on bounded fourth moment of the covariates. Using Lemma 14 for bounding the fourth moment of the noise variables, we have with probability at least $1-\delta$,

$$\mathbb{E}_{\mathcal{D}_c}\left[(\langle x,\theta\rangle - y)^4\right] \leq 8\cdot\left(C_{x,4}\|\widetilde{\Sigma}_x\|_2^2\|\Delta_\theta\|_2^4 + 3\sigma^4 + \frac{C_{z,4}\sigma^4}{\sqrt{n}}\log^2\left(\frac{e^2}{\delta}\right)\right). \tag{30}$$

Substituting the bounds obtained in Equations (29) and (30) in Equation (28), along with an application of triangle inequality, we have that with probability at least $1-\delta$,

$$\begin{aligned}\|\Sigma_\theta\|_2 \leq \underbrace{\sqrt{C_{x,4}}\|\widetilde{\Sigma}_x\|_2 + 2\sqrt{8C_{x,4}}\cdot\|\widetilde{\Sigma}_x\|_2\cdot\|\theta^* - \theta_{\text{reg}}\|_2^2 + \frac{(8C_{z,4})^{\frac{1}{2}}\cdot\sigma^2}{n^{\frac{1}{4}}}\log\left(\frac{e^2}{\delta}\right) + \sqrt{24}\sigma^2}_{C_{\Sigma,2}} \\ + \underbrace{2\sqrt{8C_{x,4}}\cdot\|\widetilde{\Sigma}_x\|_2}_{C_{\Sigma,1}}\cdot\|\theta - \theta_{\text{reg}}\|_2^2.\end{aligned} \tag{31}$$

One can now use the above bounds in conjunction with the values of $\bar{L}$ and $\bar{m}$ from Equation (27) to obtain the final result. ■

The following lemma obtains a concentration bound for the fourth moment of a Guassian random variable and can be obtained by appropriate instantiation of the Hypercontractivity Concentration Inequality (Theorem 1.9) by Schudy and Sviridenko [SS12].

**Lemma 14 (Concentration Bound for Gaussian Fourth Moment)** *Let $z_1, z_2, \ldots, z_n$ be i.i.d. random variable sampled from $\mathcal{N}(0,1)$. Then, there exists a universal constant $C_{z,r}$ such that for any $\epsilon > 0$, we have that,*

$$\Pr\left(\left|\frac{1}{n}\sum_{1=1}^{n} z_i^4 - \mathbb{E}[z^4]\right| \geq \epsilon\right) \leq e^2\exp\left(\frac{-n^{\frac{1}{4}}\epsilon^{\frac{1}{2}}}{C_{z,4}\cdot\sigma^2}\right).$$