RobULA: Efficient Sampling for Robust Bayesian Inference

Kush Bhatia [1] **Yi-An Ma** [1] **Anca D. Dragan** [1] **Michael I. Jordan** [2 1] **Peter L. Bartlett** [2 1]

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 RobULA, a robust variant of the Unadjusted Langevin Algorithm (ULA), and provide a finite-sample analysis of its sampling distribution. The key to our approach is combining a gradient-based sampling approach with a robust mean gradient estimator. In particular, we show that after $T = \tilde{\mathcal{O}}(d/\varepsilon_{\mathsf{acc}})$ iterations, we can sample from p_T such that $\text{dist}(p_T, p^*) \leq \varepsilon_{\mathsf{acc}} + \tilde{\mathcal{O}}(\epsilon)$, where ϵ is the fraction of corruptions.

1. Introduction

Robustness has raised considerable interest both in the Bayesian (DF61; BMP+94) and frequentist context (Tuk60; Hub64; Hub73a) since being introduced by George Box in 1953 (Box53). It captures the insensitivity of inferential procedures to the presence of outliers in collected observations and misspecifications in the modelling assumptions. The Bayesian approach has focused on capturing the complete behavior of observed data (including the anomalies) through robust model and prior assumptions. The frequentist approach, on the other hand, develops robust estimators to identify and guard against outliers in the observations. 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). However, Augustus Kong (Kon87) deemed a complete Bayesian approach via explicitly formulating model misspecifications infeasible in its full generality, by showing that it is computationally expensive to propagate beliefs even in finite spaces. In light of this, most existing approaches have been unable to provide a finite-sample Bayesian robustness theory. Huber goes on to ask "Why there is no finite sample Bayesian robustness theory?" (Hub11, Section 15.7).

This paper attempts to address this issue by providing a formal framework for studying Bayesian robustness theory as well as proposing a robust inferential procedure with finite-sample guarantees. In order to overcome issues of computational infeasibility, we take a different approach from the full Bayesian procedure and refrain from modelling outlier distributions 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, often termed an ϵ-contamination model, was first coined by Huber in 1964 (Hub64) and has been the subject of recent study of computationally efficient robust estimation in the frequentist setup (DKK+16; LRV16; PSBR18) (Such a model was also sought for in the Bayesian context by Thomas Richardson in (Jor11, Point 5)).

Given data corrupted in this way, our goal is to sample from the *clean* posterior distribution p^*: the posterior distribution conditioning only on the clean data. Our key idea is to leverage a well-established robust mean estimation approach in the context of gradient-based sampling techniques: by computing robust mean gradients, we generate samples robustly. We propose a Markov Chain Monte Carlo (MCMC) algorithm called RobULA which is a robust variant of the Unadjusted Langevin Algorithm (ULA). The ULA algorithm and its variants have been used for efficient large-scale Bayesian posterior sampling (WT11) and their convergence analysis has been a recent topic of interest (Dal17; CB18; DCWY18); 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 RobULA, the iterate θ_T has a distribution p_T such that $\text{dist}(p_T, p^*) \leq \varepsilon_{\mathsf{acc}} + \tilde{\mathcal{O}}(\epsilon)$, where ϵ is the fraction of corrupted points in the dataset.

[1]Department of Electrical Engineering and Computer Science, University of California, Berkeley [2]Department of Statistics, University of California, Berkeley. Correspondence to: Kush Bhatia <kush@cs.berkeley.edu>, Yian Ma <yianma@berkeley.edu>.

Presented at the ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning.

2. Related Work

2.1. Robust Statistical Procedures

There has been a huge line of work underlying the study of robust estimation procedures and outlier detection in statistics (Hub73a; Box53; DF61). In the frequentist parameter estimation setting, the most commonly studied corruption model is the Huber's ϵ-contamination model. Under the adversarial setup, many recent works are devoted to developing computationally efficient problem dependent estimators for mean estimation (LRV16; DKK+16) and linear regression (KKM18; BJK15; BJKK17; SBRJ19) as well as for general risk minimization problems (PSBR18; DKK+18). The most relevant to our setup are (PSBR18) and (DKK+18) which utilize the robust mean estimators of (LRV16; DKK+16) to robustify gradient based procedures for providing guarantees for empirical risk minimization.

The work on robustness in the Bayesian framework has mostly focussed on developing robust models and priors. An important line of work has focussed on the sensitivity of the posterior distribution to various choices of the priors (BMP+94; MSLD17; MD18) which suggest the use of non-informative priors. These methods are orthogonal to the ones considered in the paper and do not aim to robustify inferential procedures against corruptions in the observed dataset.

2.2. Sampling Methods

There have been various zeroth-order (BRS93; LS93; MT+96) and first order methods (RT+96; Erm75; N+11) proposed for sampling from distributions over continuous spaces; our focus in this paper is on the overdamped Langevin MCMC which was first proposed by (Erm75) in the context of molecular dynamics applications. Its non-asymptotic convergence (in total variation distance) was first studied by (Dal17) for log-smooth and log-strongly concave distributions. Cheng and Bartlett (CB18) then extended the analysis to obtain similar convergence result with the error measured in KL-divergence.

3. Problem Setup

Bayesian models are used in statistics to capture uncertainty around the unknown model parameters $\theta \in \mathbb{R}^d$. Given access to a dataset $\mathcal{D} = \{z_1, z_2, \ldots, z_n\}$ consisting of n data points, bayesian modelling typically requires the specification of: a) *Prior Distribution*, $p_\theta(\theta|\alpha)$, on the model parameters θ with α denoting the set of hyperparameters and b) *Likelihood*, $p(z|\theta)$, which describes the likelihood of a datapoint z being sampled from the model described by θ. Given these distributions, the posterior distribution over the parameters, $p(\theta|\mathcal{D}, \alpha)$, is then computed as:

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

One important goal of Bayesian inference is to generate a sequence of samples $\{\theta_t\}_{t=1}^T$ such that the distribution of θ_t is close to the true posterior distribution $p(\theta|\mathcal{D}, \alpha)$.

We consider the ϵ-contamination model introduced by Huber in (Hub64) and let the collection of n datapoints $\mathcal{D}$ be sampled such that each z_i is sampled from the following mixture distribution:

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

where P denotes the true underlying sampling distribution while Q is any arbitrary distribution. A Dataset $\mathcal{D}$ drawn from such a mixture distribution has each data point z_i adversarially corrupted with probability ϵ. We denote by $\mathcal{D}_c$ the subset of datapoints in $\mathcal{D}$ sampled from the true distribution P and similarly by $\mathcal{D}_a$, the subset of data sampled from Q. Given datapoints $\mathcal{D} = \mathcal{D}_c \cup \mathcal{D}_a$, the likelihood function $p(z|\theta)$ and the prior $p_\theta(\theta|\alpha)$, the objective of Robust Bayesian Inference is to obtain sample parameters θ^c from the *clean* posterior distribution $p(\theta|\mathcal{D}_c, \alpha)$ given by $p(\theta|\mathcal{D}_c, \alpha) \propto p_\theta(\theta|\alpha) \cdot \prod_{i \in \mathcal{D}_c} (z_i|\theta)$.

We now proceed to present our algorithm, RobULA, for robustly generating the sequence of samples $\{\theta_t^c\}_{t=1}^T$ and provide theoretical guarantees on its convergence properties. The main idea is to use gradient-based iterative sampling methods, and leverage a robust mean estimation procedure to compute a robust estimate of the gradient at each iteration.

Algorithm 1: RobULA: Robust Unadjusted Langevin Algorithm

Input: Dataset $\mathcal{D}$, step-size sequence η, initial covariance scaling β, timesteps T, prior distribution $p_\theta(\theta|\alpha)$, hyperparameters α, likelihood function $p(z|\theta)$, corruption level ϵ

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

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

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

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

$\theta_{k,\eta} = \theta_{k-1,\eta} - \eta_k(n \cdot \widehat{\nabla} U_\theta$
$-\nabla \log(p_\theta(\theta_{k-1,\eta}|\alpha)) + \sqrt{2\eta_k}\xi_k$

Output: Iterates $\{\theta_k\}$

4. RobULA: Robust Unadjusted Langevin Algorithm

In this section, we introduce our proposed algorithm, RobULA (Algorithm 1), for the robust posterior inference prob-

lem defined in Section 3, which is a simple modification of the ULA algorithm. In the sequel, we first briefly introduce ULA and robust gradient estimation and then proceed to describe RobULA.

4.1. Unadjusted Langevin Algorithm

In this section, we describe the Unadjusted Langevin Algorithm (ULA), for sampling from probability distributions over continuous spaces $\mathbb{R}^d$. Focusing on the space of parameters, we first rewrite the posterior distribution as:

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

where $g_i(\theta) := -\log(p(z_i|\theta))$ is the negative log-likelihood corresponding to the i^{th} datapoint. We further denote $f(\theta; \mathcal{D}) = \sum_{i\in\mathcal{D}} g_i(\theta) - \log(p_\theta(\theta|\alpha))$ so that $\exp(-f(\theta; \mathcal{D}))$. The ULA can then be used to sample from such posteriors using the updates:

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

where η and ξ_{t+1} are respectively the step-size sequence and independent Gaussian noise.

The Markov chain in Equation (2) 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

$$\mathrm{d}\theta_t = -\nabla f(\theta_t; \mathcal{D})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t , \quad t \geq 0, \quad (3)$$

where $\{B_t\}_{t\geq 0}$ represents the 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, *i.e.*, the iterates of the ULA algorithm converge in distribution to the true sampling distribution $p(\theta|\mathcal{D}, \alpha)$.

4.2. Robust Gradient Estimation

Algorithm 2 describes the 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 ϵ of them are arbitrarily corrupted. At a high level, Algorithm 2 works as follows: it projects the datapoints onto the top $d/2$ principal components of the sample covariance matrix, removes datapoints which are guaranteed to be outliers and then recursively calls itself restricted to the $d/2$-dimensional subspace. Algorithm 1 then scales this gradient estimate by the number of samples n, to obtain a robust estimate of gradients of the sum of the sample likelihoods $\sum_{i=1}^{n} \nabla g_i(\theta)$.

Note that the model described in Section 3 assumes that each datapoint z is sampled i.i.d. from the mixture distribution $(1-\epsilon)P + \epsilon Q$ where P represents the true sampling 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 ϵ_n in the sampled dataset $\mathcal{D}$ satisfy,

$$\epsilon - \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)} \leq \epsilon_n \leq \epsilon + \underbrace{\sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}}_{e_n} . \quad (4)$$

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, that is, $\left\|\widehat{\nabla} U_\theta - \frac{1}{|\mathcal{D}_c|}\sum_{i\in\mathcal{D}_c}\nabla g_i(\theta)\right\|_2$, *uniformly* for any value of the iterate θ 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 P_θ denote the corresponding distribution over $\nabla g_i(\theta)$ with mean given by ∇U_θ, covariance Σ_θ and fourth moment constant given by C_4. There exists a positive constant $C_1 > 0$ for which the robust mean estimator when instantiated with the contamination level $\gamma := \epsilon + e_n$, with probability $1-\delta$, returns an estimate $\widehat{\nabla} U_\theta$ such that for all $\theta \in \mathbb{R}^d$, we have that,*

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

RobULA Algorithm. The proposed RobULA is a simple modification of the ULA algorithm, described in Section 4.1, where in each iteration instead of using the complete set of datapoints for computing the gradient, we contruct a *robust* estimator of the gradient and update the parameter using this estimate. This robust estimator ensures that the outlier datapoints do not exert too much influence on the gradient and allow RobULA to obtain samples from a distribution close to the *clean* posterior distribution.

4.3. Convergence Analysis

In this section, we look at the convergence guarantee for the proposed algorithm RobULA. For ease of notation, we denote by $f(\theta; \mathcal{D}) = \sum_{i\in\mathcal{D}} g_i(\theta) - \log(p_\theta(\theta|\alpha))$ and similarly the clean and corrupted versions of the function $f(\theta; \mathcal{D}_c)$ and $f(\theta; \mathcal{D}_a)$ by replacing the dataset appropriately. The objective of the robust bayesian posterior inference problem is them to obtain samples from the clean posterior distribution given by $p^*(\theta|\mathcal{D}, \alpha) \propto \exp(-f(\theta; \mathcal{D}_c))$.

For clarity of exposition, we drop the dependence of the posterior distribution on the dataset $\mathcal{D}$ as well as the hyper-parameters α and denote by $f(\theta) := f(\theta; \mathcal{D}_c)$.

We now proceed to state the assumptions that we require the function $f(\theta)$ to satisy:

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

Assumption 2 (Strong convexity). The function $f(\theta)$ is m-strongly convex for all $\theta \in \mathbb{R}^d$. That is, $mI \preceq \nabla^2 f(\theta)$, $\forall \theta \in \mathbb{R}^d$. 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 sampling and optimization literature. In addition to the assumptions, we define the average Lipschitz constant $\bar{L} = L/n$ and the average strong convexity of f as $\bar{m} = m/n$. With this setup in place, we now state our main theorem concerning the convergence guarantees for RobULA.

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*

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

where Σ_θ is the covariance of uniform distribution on $\nabla g_i(\theta)$ induced by the clean dataset $\mathcal{D}_c$, $\tilde{\theta}$ satisfies $\nabla F(\tilde{\theta}) = 0$ and ϵ is the fraction of corrupted points. Then the iterates of RobULA, when initialized with $\theta_0 \sim \mathcal{N}\left(0, \frac{1}{L} I_d\right)$ (with corresponding density p_0) and step-size $\eta \leq \frac{1}{n\bar{L}}$ ($h := n\eta$), satisfy:

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

where $p_{k\eta}$ represents the distribution of the iterate $\theta_{k,\eta}$ for any $\epsilon \leq \frac{\bar{m}^2}{4 C_R C_{\Sigma,1} \log d}$.

Discussion. A few comments are in order. First observe that the error term consists of three different components: 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) captures the dependence on the fraction of corrupted points ϵ and vanishes as ϵ goes to 0, and, c) term (III) is a discretization error term.

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

$$\eta = \mathcal{O}\left(\frac{\varepsilon_{\mathsf{acc}}}{n\kappa\bar{L}d}\right) \quad \text{and} \quad T = \mathcal{O}\left(\frac{1}{m\eta} \log\left(\frac{\mathrm{KL}(p_0 \parallel p^*)}{m\varepsilon_{\mathsf{acc}}}\right)\right),$$

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 Appendix B, for typical problems like Bayesian linear regression and Bayesian mean estimation, the average strong convexity parameter $\bar{m}$ scales independent 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 selected to arbitrarily small values which would result in a corresponding increase in the number of timesteps, 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 several results in the frequentist estimation setup (BJK15; DKK+16; LRV16; PSBR18) which show that such inconsistency is a result of the adversarial corruptions and in general cannot be avoided.

5. Conclusion and Future Work

We study the problem of robustness to adversarial outliers in the Bayesian framework and propose RobULA, which is a robust extension of the classical Unadjusted Langevin Algorithm MCMC algorithm. We obtain non-asymptotic convergence guarantees for RobULA. We identify multiple directions for future work. On the statistical side, it would be interesting to extend the robustness guarantees of RobULA for statistical models which do not fall under the ambit of current assumptions, for example, non-convex likelihood functions. On the computational side, an important question to understand is whether one can accelerate the convergence of RobULA in the presence of outliers.

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, M Jesús Bayarri, José M Bernardo, Juan A Cano, Julián De la Horra, Jacinto Martín, David Ríos-Insúa, Bruno Betrò, 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.

[BRS93] Claude JP Bélisle, H Edwin Romeijn, and Robert L Smith. Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2):255–266, 1993.

[CB18] Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory*, pages 186–211, 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 *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.

[DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

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

[HCL+] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification.

[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. *Bull. Internat. Statist. Inst*, 45(4):181–191, 1973.

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

[Jor11] M Jordan. What are the open problems in bayesian statistics. *The ISBA Bulletin*, 18(1):568, 2011.

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

[Kon87] Chung Tung Augustine Kong. Multivariate belief functions and graphical models. 1987.

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

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

[MD18] Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, (just-accepted):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.

[MT+96] Kerrie L Mengersen, Richard L Tweedie, et al. 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 *2012 International Conference for Internet Technology and Secured Transactions*, pages 492–497. IEEE, 2012.

[N+11] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

[Oks03] B. 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.

[Pav14] G. A. Pavliotis. *Stochastic Processes and Applications*. Springer, 2014.

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

[RT+96] Gareth O Roberts, Richard L Tweedie, et al. 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.

[SS12] Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 437–446. Society for Industrial and Applied Mathematics, 2012.

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

[WT11] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.