DISTRIBUTIONALLY ROBUST POSTERIOR SAMPLING - A VARIATIONAL BAYES APPROACH

Bohan Wu[1,*], Bennett Zhu[1,*], David M. Blei[1,2]
[1]Department of Statistics, Columbia University
[2]Department of Computer Science, Columbia University [*]

ABSTRACT

We study the problem of robust posterior inference when observed data are subject to adversarial contamination, such as outliers and distributional shifts. We introduce *Distributionally Robust Variational Bayes (DRVB)*, a robust posterior sampling method based on solving a minimax variational Bayes problem over Wasserstein ambiguity sets. Computationally, our approach leverages gradient flows on probability spaces, where the choice of geometry is crucial for addressing different forms of adversarial contamination. We design and analyze the DRVB algorithm based on Wasserstein, Fisher-Rao, and hybrid Wasserstein-Fisher-Rao flows, highlighting their respective strengths in handling outliers, distribution shift and mixed global-local contamination. Our theoretical results establish robustness guarantees and polynomial-time convergence of each discretized gradient flow to its stationary measure. Empirical results show that DRVB outperforms the naive Langevin Monte Carlo (LMC) in generating robust posterior samples across various adversarial contamination settings.

1 INTRODUCTION

Variational Bayes is a foundational technique in modern representation learning and deep generative modeling, underpinning applications ranging from topic modeling (Blei et al., 2003) to variational autoencoders (VAE) (Kingma & Welling, 2014) and diffusion models (Song et al., 2020). However, variational inference (VI) is highly sensitive to adversarial shifts in the data distribution, and these failures can propagate throughout the probabilistic modeling pipeline, affecting both inference and generation.

Common types of adversarial contamination include:

(i) *Distributional shift*: The data distribution at deployment may differ from the one observed during training. This occurs in various settings, such as domain adaptation, where models are expected to generalize to slightly different target distributions, or in adversarial attacks, where data are manipulated specific errors—for instance, manipulating an image generation model to produce inappropriate content.

(ii) *Data contamination*: Many datasets contain outliers, measurement errors, or noise injected by privacy-preserving mechanisms (Dwork, 2006; Dwork et al., 2006a;b). As a result, the empirical distribution $\hat{\mathrm{p}}_n$ is a perturbed version of p_0.

While outlier contamination has been extensively studied in classical robust statistics (Box, 1953; Tukey, 1960; Huber, 1964), it is largely unexplored in Bayesian statsitics. Existing robust Bayesian methods primarily focus on addressing model misspecification rather than adversarial contamination. Examples include the coarsened posterior (Miller & Dunson, 2018; Bernton et al., 2019), the Wasserstein barycenter posterior (Minsker et al., 2017), and variational posteriors based on Rényi divergence (Knoblauch et al., 2022). For Huber contamination, (Bhatia et al., 2023) introduces a truncated Langevin Monte Carlo algorithm that provably outputs robust posterior mean estimates.

A popular framework to deal with data uncertainty problems is distributed robust optimization (DRO) (Blanchet et al., 2019; Duchi & Namkoong, 2021; Blanchet et al., 2024). DRO casts the

[*] Equal contribution

problem of learning under adversarial attacks as a minimax game. It has been used as a certificate of adversarial robustness for methods based on empirical risk minimization, such as linear regression or neural networks (Shafieezadeh Abadeh et al., 2015; Sagawa et al., 2019; Sinha et al., 2017). An example of DRO is the Wasserstein DRO which considers a set of pertubations based on the Wasserstein distance Nietert et al. (2022; 2023; 2024).

In this paper, we introduce *Distributionally Robust Variational Bayes (DRVB)*, a unified approach to robust posterior sampling under three types of contamination studied in recent literature (Pittas & Pensia, 2024): global contamination (Model 1), local contamination (Model 2), and mixed global-local contamination (Model 3)[1]. DRVB addresses each contamination setting by formulating a distributionally robust version of the variational Bayes problem under a suitable ambiguity set and solving it using gradient flows over the space of probability measures.

The choice of optimization geometry is crucial for effectively targeting different types of contamination. While all gradient-flow-based algorithms improve robustness compared to naive Langevin Monte Carlo, we show that the Wasserstein-Fisher-Rao (WFR) gradient flow—which interpolates between the Wasserstein and Fisher-Rao geometries—consistently outperforms both individual flows and standard Langevin Monte Carlo in the presence of mixed global and local contamination (3). Such a gain comes with a computational cost, as we prove that the WFR gradient flow algorithm takes longer to converge.

2 A Variational Representation of the Bayes Posterior

Let $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ be a global latent variable and let $\boldsymbol{x} = x_{1:n}$ denote n observations drawn from a population distribution p_0. Consider the following probabilistic model:

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \mathrm{p}(x_i \mid \boldsymbol{\theta}) \mathrm{p}(\boldsymbol{\theta}).$$

The log-joint probability decomposes additively as $\log \mathrm{p}(x, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(x_i, \boldsymbol{\theta})$, where $\ell(x_i, \boldsymbol{\theta}) := \log \mathrm{p}(x_i \mid \boldsymbol{\theta}) + \log \mathrm{p}(\boldsymbol{\theta})$.

Suppose that the posterior $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x})$ has a Lebesgue density and a finite second moment, namely $\mathbb{E}\left[\|\boldsymbol{\theta}\|^2 \mid \boldsymbol{x}\right] < \infty$. Then the posterior is the unique solution to the *variational Bayes* problem (Blei et al., 2017; Knoblauch et al., 2022):

$$\mathrm{q}^*(\boldsymbol{\theta}) = \underset{\mathrm{q} \in \mathscr{P}_{2,ac}(\Theta)}{\arg\min} \left\{ -\mathbb{E}_{\mathrm{q}(\boldsymbol{\theta})} \left[\sum_{i=1}^{n} \ell(x_i, \boldsymbol{\theta}) \right] + \mathbb{H}(\mathrm{q}) \right\}. \tag{1}$$

The problem above is equivalent to minimizing the Kullback-Leibler (KL) divergence between some distribution q and the exact posterior:

$$\mathrm{q}^*(\boldsymbol{\theta}) = \underset{\mathrm{q} \in \mathscr{P}_{2,ac}(\Theta)}{\arg\min} \; D_{\mathrm{KL}}(\mathrm{q} \,\|\, \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x})). \tag{2}$$

Since the posterior is in the set $\mathscr{P}_{2,ac}(\Theta)$, the minimizer is achieved at the posterior $\mathrm{q}^*(\boldsymbol{\theta}) = \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x})$. But the purpose of Eq. 1 is to offer a perspective shift: it recasts the task of *posterior sampling*—that is, generating samples from $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x})$—as the problem of solving an infinite-dimensional KL-minimization problem for $\mathrm{q}^*(\boldsymbol{\theta})$. The latter enables natural extensions of the posterior sampling problem by modifying the objective to incorporate robustness constraints. Moreover, it enables us to consider geometries on the space of probability measures and the derivation of gradient dynamics for minimizing the KL functional under the chosen geometry.

3 Distributionally Robust Variational Bayes (DRVB)

Let $\hat{\mathrm{p}}_n$ denote the empirical distribution of the observed data $x_{1:n}$. Problem (1) can be rewritten as:

$$\mathrm{q}^*(\boldsymbol{\theta}) = \underset{\mathrm{q} \in \mathscr{P}_{2,ac}(\Theta)}{\arg\min} \; -n\mathbb{E}_{\mathrm{q}(\boldsymbol{\theta})} \left[\mathbb{E}_{\hat{\mathrm{p}}_n(x)} \left[\ell(x, \boldsymbol{\theta}) \right] \right] + \mathbb{H}(\mathrm{q}). \tag{3}$$

[1] Formal definitions are provided in Appendix A. At a high level, local contamination is measured under the p-Wasserstein distance, while global contamination is defined under the total variation (TV) distance.

The basic assumption of Bayesian inference is that the empirical measure $\hat{\mathrm{p}}_n$ accurately approximates p_0 so that $\mathrm{q}^*(\theta)$ captures the uncertainty of the parameter θ under the population distribution.

However, when the observed data aree contaminated, $\hat{\mathrm{p}}_n$ no longer accurately represents the population distribution p_0. This lack of robustness arises because deviations of $\hat{\mathrm{p}}_n$ from p_0 introduce an error factor of n into (3), which leads to biased posterior inference.

To deal with adversarial contamination, the *distributionally robust optimization (DRO)* framework introduces an uncertainty set to account for discrepancies between the in-sample distribution $\hat{\mathrm{p}}_n$ and the out-of-sample distribution. Motivated by the DRO approach, we propose *Distributionally Robust Variational Bayes (DRVB)*, which minimizes the worst-case variational Bayes objective within the uncertainty set:

$$\left(\mathrm{q}_\delta^*(\theta), \mathrm{p}_\delta^*(x)\right) = \underset{\mathrm{q}\in\mathscr{P}_{2,ac}(\Theta)}{\arg\min} \max_{\substack{\mathbb{W}_p(\mathrm{p},\hat{\mathrm{p}}_n)\leq\delta;\\ \mathrm{p}\in\mathscr{P}_p(\mathscr{X})}} \mathscr{F}(\mathrm{q},\mathrm{p}), \tag{P1}$$

where

$$\mathscr{F}(\mathrm{q},\mathrm{p}) := -n\mathbb{E}_{\mathrm{q}(\theta)}\left[\mathbb{E}_{\mathrm{p}(x)}\left[\ell(x,\theta)\right]\right] + \mathbb{H}(\mathrm{q}). \tag{5}$$

We refer to $\mathrm{q}_\delta^*(\theta)$ as the δ*-DR posterior*.

The DRVB problem is an infinite-dimensional minimax optimization problem and $\left(\mathrm{q}_\delta^*(\theta), \mathrm{p}_\delta^*(x)\right)$ is the Nash equilibrium. Since the constraint set $\mathscr{P}_{2,ac}(\Theta)$ is a geodesically-convex subset of $\mathscr{P}_2(\Theta)$. we could leverage Langenvin dynamics to design a sampling algorithm (Jordan et al., 1998).

4 Algorithms

Let $\mathscr{X}$ be a subset of $\mathbb{R}^p$. Consider the Lagrangian DRVB problem:

$$\mathrm{q}_\lambda^*(\theta) = \underset{\mathrm{q}\in\mathscr{P}_{2,ac}(\Theta)}{\arg\min} \max_{\mathrm{p}\in\mathscr{P}_p(\mathscr{X})} \left\{\mathbb{H}(\mathrm{q}) - n\mathbb{E}_{\mathrm{q}(\theta)}\left[\mathbb{E}_{\mathrm{p}(x)}\left[\ell(x,\theta)\right]\right] - \lambda\mathbb{W}_p^p(\mathrm{p},\hat{\mathrm{p}}_n)\right\}. \tag{6}$$

Since problem (4) is convex in p, the regularization weight λ in Eq. 6 is equivalent to the radius δ of the ambiguity set up to some monotone transformation. A large δ corresponds to a small λ, and vice versa.

To solve the min-max problem for $\mathrm{q}_\lambda^*(\theta)$, we use a *Wasserstein Gradient Descent Ascent (WGDA)* algorithm. At each iteration, the algorithm performs Wasserstein gradient descent over $\mathrm{p}(\boldsymbol{x})$ and gradient ascent over $\mathrm{q}(\theta)$.

Take $\Pi(\mathrm{p},\hat{\mathrm{p}}_n)$ to be the set of all coupling of $\mathrm{p},\hat{\mathrm{p}}_n$. We can write the adversarial cost as a semi-discrete optimal transport problem,

$$\mathbb{W}_p^p(\mathrm{p},\hat{\mathrm{p}}_n) = \inf_{\pi\in\Pi(\mathrm{p},\hat{\mathrm{p}}_n)} \int_{\mathscr{X}\times\mathscr{X}} \|x-y\|^p d\pi(x,y).$$

The Wasserstein distance finds the optimal π in the set of couplings. But the cost could be defined for general $\pi\in\mathscr{P}_p(\mathscr{X}\times\mathscr{X})$, given by

$$\mathscr{C}_p^p(\pi) := \int_{\mathscr{X}\times\mathscr{X}} \|x-y\|^p d\pi(x,y).$$

We use π_x,π_y to denote the first and second marginals of π. In our context, π_x,π_y corresponds to the perturbed and empirical distributions, respectively.

Given q and π, define the objective:

$$\mathscr{U}(\mathrm{q},\pi) := \mathbb{H}(\mathrm{q}) - n\mathbb{E}_{\mathrm{q}(\theta)}\left[\mathbb{E}_{\pi_x(x)}\left[\ell(x,\theta)\right]\right] - \lambda\mathscr{C}_p^p(\pi).$$

Then Eq. 6 is equivalent to the following reparametrized problem

$$\mathrm{q}_\lambda^*(\theta) = \underset{\mathrm{q}\in\mathscr{P}_{2,ac}(\Theta)}{\arg\min} \max_{\pi\in\mathscr{P}_p(\mathscr{X}\times\mathscr{X});\pi_y=\hat{\mathrm{p}}_n} \mathscr{U}(\mathrm{q},\pi). \tag{7}$$

It is easy to see that the two minimizers agree in Eq. 6 and Eq. 7.

4.1 Wasserstein Gradient Descent Ascent

Let $s_\theta(x,\theta)$ and $s_x(x,\theta)$ be the partial gradient of the log joint $\ell(x,\theta)$ with respect to θ and x, respectively.

We firs take the Wassestein gradient of $\mathscr{U}(\mathrm{q},\pi)$ with respect to q, given by the gradient of the first variation of $\mathscr{U}$ in the direction of q.

$$\nabla_{\mathbb{W}_2,\mathrm{q}}\mathscr{U}(\mathrm{q},\pi) = \nabla\left(\frac{d}{d\varepsilon}\mathscr{U}\left(\mathrm{q}+\varepsilon(\mathrm{q}'-\mathrm{q}),\pi\right)\right) = -n\mathbb{E}_{\pi_x(x)}\left[s_\theta(x,\theta)\right] + \nabla\log\mathrm{q}(\theta). \tag{8}$$

For a given π, let $(\mathrm{q}_t^\pi)_{t\geq 0}$ be the gradient flow that evolves according to Eq. 8. Let $\theta_t^\pi \sim \mathrm{q}_t^\pi$ be a random variable at time t.

Using the well-known result of (Jordan et al., 1998), $(\theta_t^\pi)_{t\geq 0}$ follows the *Langevin dynamics*:

$$d\theta_t^\pi = -n\mathbb{E}_{\pi_x(x)}\left[s_\theta(x,\theta_t^\pi)\right]dt + \sqrt{2}dB_t. \tag{9}$$

At the level of trajectory $(\theta_t^\pi)_{t\geq 0}$, the Langevin dynamic is a stocahstic differential equation (SDE) but at the level of measures $(\mathrm{q}_t^\pi)_{t\geq 0}$, it is a gradient flow under the Wasserstein geometry.

Now we differentiate $\mathscr{U}(\mathrm{q},\pi)$ with respect to π. The first step is to rewrite $\mathscr{U}(\mathrm{q},\pi)$ as follows,

$$\mathscr{U}(\mathrm{q},\pi) = \mathbb{E}_{\pi(x,y)}\left[-n\mathbb{E}_{\mathrm{q}(\theta)}\left[\ell(x,\theta)\right] - \lambda\|x-y\|^p\right] + \mathbb{H}(\mathrm{q}).$$

This shows that $\mathscr{U}(\mathrm{q},\pi)$ depends linearly on the optimization variable π. The Wasserstein gradient for π is the gradient of first variation:

$$\nabla_{\mathbb{W}_2,\pi}\mathscr{U}(\mathrm{q},\pi) = \nabla_{x,y}\left(-n\mathbb{E}_{\mathrm{q}(\theta)}\left[\ell(x,\theta)\right] - \lambda\|x-y\|^p\right) = \begin{bmatrix} -n\mathbb{E}_{\mathrm{q}(\theta)}\left[s_x(x,\theta)\right] - \lambda\nabla_x(\|x-y\|^p) \\ -\lambda\nabla_y(\|x-y\|^p) \end{bmatrix}. \tag{10}$$

where $\nabla_x(\|x-y\|^p) = p\|x-y\|^{p-1}\nabla_x(\|x-y\|) = p\|x-y\|^{p-1}\|x-y\|^{-1}x = p\|x-y\|^{p-2}x$.

The gradient flow of π evolves a trajectory $(x_t^\mathrm{q}, y_t^\mathrm{q})_{t\geq 0}$ according to a Euclidean gradient ascent flow:

$$dx_t^\mathrm{q} = -n\mathbb{E}_{\mathrm{q}(\theta)}\left[s_x\left(x_t^\mathrm{q},\theta\right)\right]dt - \lambda\nabla_x(\|x_t^\mathrm{q}-y_t^\mathrm{q}\|^p)dt, \tag{11}$$

and

$$dy_t^\mathrm{q} = -\lambda\nabla(\|x_t^\mathrm{q}-y_t^\mathrm{q}\|^p)dt. \tag{12}$$

The joint gradient for DRVB glues together the Langevin dynamics $(\theta_t^\pi)_{t\geq 0}$ and the bivariate flow $(x_t^\mathrm{q}, y_t^\mathrm{q})_{t\geq 0}$. For $\theta_t \sim \mathrm{q}_t$ and $(x_t,y_t)\sim\pi_t$, we obtain:

$$\begin{aligned} d\theta_t &= n\mathbb{E}_{\pi_{t,x}(x)}\left[s_\theta(x,\theta_t)\right]dt + \sqrt{2}dB_t, \\ dx_t &= -n\mathbb{E}_{\mathrm{q}_t(\theta)}\left[s_x(x_t,\theta)\right]dt - \lambda\nabla(\|x_t-y_t\|^p)dt, \\ dy_t &= -\lambda\nabla(\|x_t-y_t\|^p)dt. \end{aligned} \tag{13}$$

Next, we project the unconstrained gradient flow (13) onto the subspace $\{\pi\in\mathscr{P}_p(\mathscr{X}\times\mathscr{X}) : \pi_y = \hat{\mathrm{p}}_n\}$. In particular, we enforce $y_0\sim\pi_0 := \hat{\mathrm{p}}_n$, thus $dy_t = 0$. This results in the following flow:

$$\begin{aligned} &\text{(Sampling)} \quad d\theta_t = n\mathbb{E}_{\pi_{t,x}(x)}\left[s_\theta(x,\theta_t)\right]dt + \sqrt{2}dB_t, \\ &\text{(Adversary)} \quad dx_t = -n\mathbb{E}_{\mathrm{q}_t(\theta)}\left[s_x(x_t,\theta)\right]dt - \lambda\nabla(\|x_t-y_0\|^p)I\{(x_t,y_0)\in\mathrm{supp}(\pi_t)\}dt. \end{aligned} \tag{14}$$

We refer the first flow in Eq. 14 as *θ-flow* and the second flow as *x-flow*. We briefly check the joint flow in special cases. As $\lambda\to\infty$, x-flow enforces the hard constraint $x_t = y_0$ and the θ-flow reduces to the Langevin dynamics for posterior sampling. When $p=2$, x-flow becomes $dx_t = -n\mathbb{E}_{\mathrm{q}_t(\theta)}\left[s_x(x_t,\theta)\right]dt - \lambda p(x_t-y_0)I\{(x_t,y_0)\in\mathrm{supp}(\pi_t)\}dt$.

Discretization and implementation. Our algorithm is a time-discretized version of the joint flow in Eq. 14. To discretize the gradient flow, we use the Euler-Maruyama scheme.

The discretized θ-flow becomes the Lagevin Monte Carlo algorithm. For a step size η_k, it updates

$$\theta_{(k+1)\eta_k} = \theta_{k\eta_k} + \eta_k n \nabla \mathbb{E}_{\pi_{k\eta_k,x}(x)} \left[s_\theta \left(x, \theta_{k\eta_k} \right) \right] + \sqrt{2\eta_k}\xi, \quad \xi \sim \mathcal{N}\left(0, I_d\right). \tag{15}$$

The discretized x-flow is the vanilla gradient ascent algorithm with step size η_k,

$$x_{(k+1)\eta_k} = x_{k\eta_k} - \eta_k n \mathbb{E}_{\mathrm{q}_{k\eta_k}(\theta)} \left[s_x \left(x_{k\eta_k}, \theta \right) \right] dt - \eta_k \lambda p \| x_{k\eta_k} - y_0 \|_p^{p-2} (x_{k\eta_k} - y_0). \tag{16}$$

The detailed algorithm is described in Algorithm 1.

```
Algorithm 1: Wasserstein Gradient Descent Ascent
Input: Initial particles $\theta_0^1, \cdots, \theta_0^m$, Data $x_0^1, \cdots, x_0^n$, Terminal time $T$, Step sizes $\eta_{0:(T-1)}$
for $k = 0$ to $T-1$ do
    Set $\mathbf{v}_k = 0$;
    Use $n_k$ Monte Carlo samples to estimate;
        $\mathbb{E}_{\pi_{k,x}(x)} \left[ s_\theta \left( x, \theta_k \right) \right] \approx \frac{1}{n_k} \sum_{i=1}^{n_k} s_\theta \left( x^i, \theta_k \right) \quad \text{s.t. } x^1, \cdots, x^{n_k} \overset{iid}{\sim} \frac{1}{n} \sum_{i=1}^{n} \delta_{x^i_{k-1 \vee 0}}.$
    Compute the corresponding estimator $\mathbf{v}_k$;
    Update $\theta_{k+1}^j = \theta_k^j + \frac{\eta_k}{\lambda} n \mathbb{E}_{\pi_{k,x}(x)} \left[ s_\theta \left( x, \theta_k^j \right) \right] + \sqrt{2\eta_k} \xi^j$, where $\xi^j \sim \mathcal{N}\left(0, I_d\right)$;
    Use $m_k$ Monte Carlo samples to estimate;
        $\mathbb{E}_{\mathrm{q}_k(\theta)} \left[ s_x \left( x_k, \theta \right) \right] \approx \frac{1}{m_k} \sum_{j=1}^{m_k} s_x \left( x_k, \theta^j \right) \quad \text{s.t. } \theta^1, \cdots, \theta^{m_k} \overset{iid}{\sim} \frac{1}{m} \sum_{j=1}^{m} \delta_{\theta_k^j}.$
    Run Sinkhorn algorithm to compute $\pi_k^*$ is the optimal coupling between $\hat{\mathrm{p}}_n$ and $\mathrm{p}_k$.
    $\nabla_{\mathbb{W},\mathrm{p}} \mathscr{U}(q_k, p_k)(x) = \nabla \delta \mathscr{U}_{\mathrm{p}}(\mathrm{q}_k, \mathrm{p}_k)(x) = -n \mathbb{E}_{\mathrm{q}_k(\theta)} \left[ s_x(x, \theta) \right] - \lambda (\nabla_x \| x - y \|^p) I \{ (x, y) \in \mathrm{supp}(\pi_k^*) \}$
    Update $x_{k+1}^j = x_k^j - \frac{\eta_k}{\lambda} \nabla_{\mathbb{W},\mathrm{p}} \mathscr{U}(q_k, p_k)(x_k^j)$;
return $\theta_T, x_T$;
```

At each iteration, we approximate the expected score in θ and x with Monte Carlo samples from the latest updates. The sequence of samples sizes m_k, n_k are tuned between 1 and m, n based on computational budget. Setting $m_k = n_k = 1$ corresponds to a full stochastic gradient approach, setting $m_k = n_k = n$ is a deterministic gradient descent ascent, setting m_n, n_k in between corresponds to mini-batch optimization.

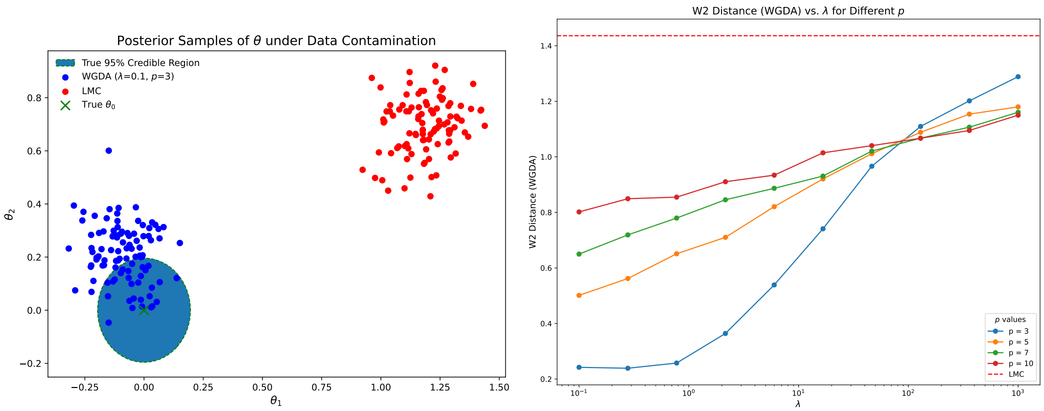
We set the stepsize according to the Robbins-Monroe sequence, i.e. η_k / λ satisfies $\sum_{k=0}^{\infty} \eta_k = \infty$ and $\eta_k^2 < \infty$. Dividing by λ offsets the large discretization error to the gradient flow induced by a large choice of λ.

5 A Toy illustration

Consider a bivariate normal model. The log-joint is given by $\ell(x, \theta) = -\frac{1}{2} \| x - \theta \|^2$ for data $x \sim \mathrm{p}_0 := \mathcal{N}(0, I_2)$.

We simulate a dataset of 100 points from the bivariate normal distribution $\mathcal{N}(0, I_2)$. We then contaminate this dataset by introducing 20 outliers. These outliers are generated from an Exponential(1) distribution, and then scaled by a factor of 10.

We run the proposed Wasserstein Gradient Descent Ascent (WGDA) and Langevin Monte Carlo (LMC) algorithms on the contaminated samples. Interestingly, the WGDA algorithm shows instability when $p = 2$ and λ is small. When $\ell(x, \theta) = -\frac{1}{2} \| x - \theta \|^2$, higher-order regularization or a large λ is necessary to ensure the problem is "concave" in the distribution of x. For instance, when $p = 2$,



(a) Scatter plot comparing the posterior samples from WGDA and LMC, under outlier contamination. The blue region represents the true 95% credible region when the data is uncontaminated.

(b) W_1 distance between WGDA posterior samples and clean posterior samples for varying λ and p.

Figure 1: Comparing posterior samples from WGDA and LMC (left) and W_2 distance between WGDA posterior samples and clean posterior samples (right).

we need $\lambda > 50$ to make the inner problem concave. We recommend setting $p > 2.5$ to allow for a relatively smaller choice of λ.

The WGDA is run with a terminal time $T = 1$, step size $\eta = 0.001$, $m_k = n_k = 10$, Wasserstein power $p = 3$ and regularization parameter $\lambda = 0.1$. As we vary the size of λ, smaller λ makes the result of WGDA more robust to outliers. The competing LMC algorithm runs with the same step size $\eta = 0.001$ as WGDA.

Figure 1a shows that WGDA produces samples that closely match the posterior credible region under the clean distribution p_0, whereas LMC produces unreliable samples due to data contamination.

Figure 1b plots the discrepancy between the posterior samples generated by WGDA and those from the target posterior. For smaller values of p, there is a more pronounced improvement in performance as λ decreases. WGDA outperforms LMC on all (p, λ) settings considered.

6 Discussion

In this manuscript, we formulate the Distributionally Robust Variational Bayes (DRVB) problem for posterior inference under adversarial contamination, introduce a Wasserstein Gradient Descent Ascent (WGDA) algorithm as a robust posterior sampling method, and demonstrate its effectiveness on a contaminated Gaussian example.

In the appendix, we present a variant of the algorithm that optimizes DRVB under the Fisher-Rao geometry, which is particularly well-suited for robust sampling under global contamination, and Wasserstein-Fisher-Rao (WFR) geometry, which addresses mixed global and local contamination.

Immediate future work involves establishing theoretical guarantees, including convergence analysis and robustness properties under various contamination settings Model 1 to 3.

Acknowledgments

We thank Nicolas Garcia Trillos for valuable discussions.

References

Espen Bernton, Pierre E. Jacob, Mathieu Gerber, and Christian P. Robert. Approximate bayesian computation with the wasserstein distance. *Journal of the Royal Statistical Society. Series B:*

Statistical Methodology, 81, 2019. ISSN 14679868. doi: 10.1111/rssb.12312.

Kush Bhatia, Yi An Ma, Anca D. Dragan, Peter L. Bartlett, and Michael I. Jordan. Bayesian robustness: A nonasymptotic viewpoint. *Journal of the American Statistical Association*, 2023. ISSN 1537274X. doi: 10.1080/01621459.2023.2174121.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37:165–191, 2019.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56, 2019. ISSN 00219002. doi: 10.1017/jpr.2019.49.

Jose Blanchet, Jiajin Li, Sirui Lin, and Xuhui Zhang. Distributionally robust optimization and robust statistics. *arXiv preprint arXiv:2401.14655*, 2024.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877, 2017.

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

José A Carrillo, Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, and Dongyi Wei. Fisher-rao gradient flow: geodesic convexity and functional inequalities. *arXiv preprint arXiv:2407.15693*, 2024.

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M Stuart. Sampling via gradient flows in the space of probability measures. *arXiv preprint arXiv:2310.03597*, 2023.

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M Stuart. Efficient, multimodal, and derivative-free bayesian inference with fisher–rao gradient flows. *Inverse Problems*, 40(12):125001, 2024.

Sinho Chewi. *Log-Concave Sampling*. draft, 2023.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49, 2021. ISSN 21688966. doi: 10.1214/20-AOS2004.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4004 LNCS, 2006a. doi: 10.1007/11761679_29.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3876 LNCS, 2006b. doi: 10.1007/11681878_14.

Zhou Fan, Leying Guan, Yandi Shen, and Yihong Wu. Gradient flows for empirical Bayes in high-dimensional linear models. 12 2023. URL https://arxiv.org/abs/2312.12708v1.

Peter J Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(4):73–101, 1964.

R Jordan, D Kinderlehrer, and F Otto. The variational formulation of the fokker-planck equation. *SIAM J. Math. Anal.*, 29:1–17, 1 1998. ISSN 0036-1410.

Diederik P. Kingma and Max Welling. Auto-encoding variational {Bayes}. 2014. doi: 10.61603/ceas.v2i1.33.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayesŕule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23:1–109, 2022.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 2018.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B Dunson. Robust and scalable Bayes via a median of subset posterior measures. *Journal of Machine Learning Research*, 18: 4488–4527, 2017. ISSN 1532-4435.

Henrique K Miyamoto, Fábio CC Meneghetti, Julianna Pinele, and Sueli IR Costa. On closed-form expressions for the fisher–rao distance. *Information Geometry*, pp. 1–44, 2024.

Sloan Nietert, Ziv Goldfeld, and Rachel Cummings. Outlier-robust optimal transport: Duality, structure, and statistical analysis. In *International Conference on Artificial Intelligence and Statistics*, pp. 11691–11719. PMLR, 2022.

Sloan Nietert, Rachel Cummings, and Ziv Goldfeld. Robust estimation under the wasserstein distance. *arXiv preprint arXiv:2302.01237*, 2023.

Sloan Nietert, Ziv Goldfeld, and Soroosh Shafiee. Outlier-robust wasserstein dro. *Advances in Neural Information Processing Systems*, 36, 2024.

Thanasis Pittas and Ankit Pensia. Optimal robust estimation under local and global corruptions: Stronger adversary and smaller error. *arXiv preprint arXiv:2410.17230*, 2024.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 55. Springer, 2015.

Soroosh Shafieezadeh Abadeh, Peyman M Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. *Advances in neural information processing systems*, 28, 2015.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Qifan Song, Yan Sun, Mao Ye, and Faming Liang. Extended stochastic gradient mcmc for large-scale Bayesian variable selection. *arXiv:2002.02919*, 2020.

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

C Villani. *Topics in optimal transportation*. American Mathematical Society, 2003.

C Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

Yuling Yan, Kaizheng Wang, and Philippe Rigollet. Learning Gaussian mixtures using the Wasserstein-fisher-rao gradient flow. *arXiv preprint arXiv:2301.01766*, 2023.

A Models of Contamination

We define three models of contamination for the analysis of our algorithms:

Model 1. *TV Contamination(global contamination)*

Let $\varepsilon \in (0,1/2)$, S be a multi-set of n points in $\mathbb{R}^d$. Consider all n-sized sets in $\mathbb{R}^d$ that differ in at most ε-fraction of points, i.e., $\mathscr{O}(S,\varepsilon) := \left\{S' \subset \mathbb{R}^d : |S'| = n \text{ and } |S' \cap S| \geq (1-\varepsilon)n\right\}$. The adversary can return any set $T \in \mathscr{O}(S,\varepsilon)$. We call points in S to be the inliers and points in $T \backslash S$ to be the outliers.

Model 2. *Wasserstein-p Contamination(local contamination)*

Let $\rho \geq 0, p \in [1,\infty), S_0 = \{x_1, \cdots, x_n\}$ be an n-sized set in $\mathbb{R}^d$. Consider an adversary that perturbs each point x_i to $\tilde{x}_i$ with the average W_1 perturbation(over n points) is at most ρ. Formally, we define:

$$\begin{aligned}\mathscr{W}_1(S_0,\rho) &:= \left\{ S = \{\tilde{x}_1, \cdots, \tilde{x}_n\} \mid \frac{1}{n}\sum_{i\in[n]} \|\tilde{x}_1 - x_i\| \leq \rho \right\}, \quad p = 1 \\ \mathscr{W}_p(S_0,\rho) &:= \left\{ S = \{\tilde{x}_1, \cdots, \tilde{x}_n\} \mid \frac{1}{n}\sum_{i\in[n]} \|\tilde{x}_1 - x_i\|^p \leq \rho \right\}, \quad p > 1\end{aligned} \tag{17}$$

The adversary returns an arbitrary set $S \in \mathscr{W}_p(S_0,\rho), p \in [1,\infty)$.

Model 3. *TV and Wasserstein Contamination(global and local contamination)*

Let $\varepsilon \in (0,1/2)$ and $\rho > 0$. Let $S_0 = \{x_1, \cdots, x_n\}$ be a set of n points in $\mathbb{R}^d$. The adversary can return an arbitrary set T such that $T \in \mathscr{O}(S,\varepsilon)$ for some $S \in \mathscr{W}_p(S_0,\rho)$.

B Types of Geometries

B.1 Wasserstein Geometry

Suppose $x \in \mathscr{X}$ and $\theta \in \Theta$. We denote $\Pi(\mathrm{p},\mathrm{q})$ as the set of couplings between measures p and q, and $\mathscr{P}(\Theta)$ as the space of probability measures over Θ, $\|\cdot\|$ as the Euclidean distance in $\mathbb{R}^d$. In this paper, we work with measures with well-defined densities and we tacitly identify the measures $\mathbb{P},\mathbb{Q}$ with the densities p,q.

For $p \geq 1$, we define the (p^{th})-Wasserstein distance as follows

$$\mathbb{W}_p(\mathrm{p},\mathrm{q}) = \left(\inf_{\pi\in\Pi(\mathrm{p},\mathrm{q})} \mathbb{E}_{\pi(x,y)} \left[\|X-Y\|^p\right] \right)^{1/p}.$$

The *Wasserstein space* of order p is defined as

$$\mathscr{P}_p(\Theta) := \left\{ \mu \in \mathscr{P}(\Theta); \int_\Theta d(\theta_0,\theta)^p \mu(d\theta) < \infty \right\},$$

where $\theta_0 \in \Theta$ is arbitrary. The $\mathbb{W}_k$ distance then defines a (finite) metric on $\mathscr{P}_p(\Theta)$.

We denote by $\mathscr{P}_{p,ac}(\Theta)$ the subspace of $\mathscr{P}_p(\Theta)$ consisting of measures with Lebesgue densities. Among them, a speical space $(\mathscr{P}_{2,ac}(\Theta),\mathbb{W}_2)$ is a metric space with pseudo-Riemannian geometry (Villani, 2009).

The subspace of $\left(\mathscr{P}_{2,ac}(\mathbb{R}^d),\mathbb{W}_2\right)$ consisting of all Gaussian distributions is known as the *Bures-Wasserstein space*, denoted as $\mathrm{BW}(\mathbb{R}^d)$ (Bhatia et al., 2019).

See (Villani, 2003; 2009; Santambrogio, 2015) for a textbook treatment of optimal transport and (Chewi, 2023) for log-concave sampling theory.

Let $\mathbb{H}(\mathrm{q}) := \mathbb{E}_\mathrm{q}[\log \mathrm{q}(\theta)]$ denote the Boltzmann entropy (Villani, 2003). A key result in optimal transport is that the relative entropy functional $\mathscr{F}(\mathrm{q})$ is α*-geodesically-convex* in the Wasserstein sense

$$\mathscr{F}(\mathrm{q}) := \mathbb{E}_\mathrm{q}[V(\theta)] + \mathbb{H}(\mathrm{q}). \tag{18}$$

provided that the function V is α-strongly convex.

We use several notions of convexity for functionals on $\mathscr{P}(\mathscr{X})$: *geodesic convexity* and *linear convexity*. Geodesic convexity is defined under geodesics under a certain geometry(for example $\mathbb{W}_2$ geometry), while linear convexity is defined with respect to mixtures of two distributions (see (Villani, 2009, Ch. 16 and 17)). Following the convention (Villani, 2003), we say a functional $\mathscr{F}(\mathrm{p})$ is *convex* if it is convex in the linear geometry and *geodesically convex* otherwise.

For any functional $\mathscr{F}(\mathrm{p})$ of a positive density p on a compact set $\mathscr{X}$, we define its *first variation* $\delta\mathscr{F}(\mathrm{p}) : \mathscr{X} \mapsto \mathbb{R}$ as the unique function, up to an additive constant, for which

$$\frac{d}{d\varepsilon}\,|_{\varepsilon=0}\, \mathscr{F}(\mathrm{p}+\varepsilon\chi) = \int_{\mathscr{X}} \delta\mathscr{F}(\mathrm{p})d\chi(x), \tag{19}$$

for every signed measure $d\chi$ satisfying that $\int_{\mathscr{X}} d\chi(x) = 0$.

The Wasserstein-2 gradient is defined as the gradient of first variation:

$$\nabla_{\mathbf{W2}}\mathscr{F}(\mathrm{p}) = \nabla\delta\mathscr{F}(\mathrm{p}). \tag{20}$$

B.2 Fisher-Rao geometry

The study of dynamics of probability density functions with respect to the Fisher-Rao geometry is a newly emerging research subject across machine learning and Bayesian statistics(Carrillo et al., 2024; Chen et al., 2023; 2024; Miyamoto et al., 2024). (Carrillo et al., 2024) studies the geodesic convexity and functional inequalities of Fisher-Rao geometry and establishes the universal exponential convergence rate of Fisher-Rao gradient flow with minimal assumptions on the target distribution via the dual gradient dominance condition. (Chen et al., 2023) studies the Fisher-Rao metric from the perspective of invariance, which is a desirable property for sampling highly anisotropic target distributions. (Miyamoto et al., 2024) surveys the available closed-form expressions for the Fisher-Rao distance of both discrete and continuous distributions. (Chen et al., 2024) applies the Fisher-Rao gradient flow to propose efficient Bayesian inference methods in the presence of multiple modes, infeasibility of gradient of density, need for repeated evaluations. Their work focuses mainly on gaussian Their work mainly focuses on Gaussian approximation and Gaussian mixture approximation.

The Fisher-Rao gradient of a functional $\mathscr{F} : \mathscr{P}(\mathbb{R}^d)$ is the first variation of $\mathscr{F}$:

$$\nabla_{\mathbf{FR}}\mathscr{F}(\mathrm{p}) = \delta\mathscr{F}(\mathrm{p}). \tag{21}$$

B.3 Wasserstein-Fisher-Rao geometry

Even though WFR metric does not have an explicit form, the WFR gradient is simply a combination of Wasserstein gradient and Fisher-Rao gradient, which makes the implementation of Wasserstein-Fisher-Rao gradient flow possible. Hybrid flows under combined Fisher-Rao and Wasserstein-2 gemetries were applied to statistical problems in (Yan et al., 2023; Fan et al., 2023). (Yan et al., 2023) first introduced Wasserstein-Fisher-Rao gradient flow to statistical problems(learning gaussian mixtures). (Fan et al., 2023) then applied the hybrid flows to Empirical Bayes by introducing another random variable.

The Wasserstein-Fisher-Rao gradient of a functional $\mathscr{F} : \mathscr{P}(\mathbb{R}^d) \mapsto \mathbb{R}$ composes the Wasserstein and Fisher-Rao gradient.

$$\nabla_{\mathbf{WFR}}\mathscr{F}(\mathrm{p}) = (\nabla_{\mathbf{W2}}\mathscr{F}(\mathrm{p}), \nabla_{\mathbf{FR}}\mathscr{F}(\mathrm{p})) \tag{22}$$

C Additional Algorithms and Simulation Results

C.1 Fisher-Rao Gradient Descent Ascent

Algorithm 2: Fisher-Rao Gradient Descent Ascent

```
Input: Initial particles $\theta_0^1, \cdots, \theta_0^m$, Data $x_0^1, \cdots, x_0^n$, Weights $\boldsymbol{w}_0^i = 1/n$ for $i \in [n]$ , Terminal
    time $T$, $\mathrm{p}_0 = \sum_{i=1}^n \boldsymbol{w}_0^j \delta_{x_0^j}$, Step size $\eta_k$, Cut-off time $L$, Cut-off size $n_r$
for $k = 0$ to $T - 1$ do
    Set $\mathbf{v}_k = 0$;
    Use $n_k$ Monte Carlo samples to estimate;
        $\mathbb{E}_{\pi_{k,x}(x)}\left[s_\theta\left(x, \theta_k\right)\right] \approx \frac{1}{n_k} \sum_{i=1}^{n_k} s_\theta\left(x^i, \theta_k\right) \quad \text{s.t. } x^1, \cdots, x^{n_k} \overset{iid}{\sim} \mathrm{p}_k = \sum_{i=1}^n \boldsymbol{w}_k^j \delta_{x_k^i}.$
    Compute the corresponding estimator $\mathbf{v}_k$;
    Update $\theta_{k+1}^j = \theta_k^j + \eta_k n \mathbb{E}_{\pi_{k,x}(x)}\left[s_\theta\left(x, \theta_k^j\right)\right] + \sqrt{2\eta_k}\xi^j$, where $\xi^j \sim \mathcal{N}\left(0, I_d\right)$;
    if $k \leq L$ then
        Use $m_k$ Monte Carlo samples to estimate;
            $\mathbb{E}_{\mathrm{q}_k(\theta)}\left[s_x\left(x_k, \theta\right)\right] \approx \frac{1}{m_k} \sum_{j=1}^{m_k} s_x\left(x_k, \theta^j\right) \quad \text{s.t. } \theta^1, \cdots, \theta^{m_k} \overset{iid}{\sim} \frac{1}{m} \sum_{j=1}^m \delta_{\theta_k^j}.$
        Run Sinkhorn to compute $\pi_k^*$, the optimal coupling between $\mathrm{p}_0$ and $\mathrm{p}_k$, and potential
          functions $\phi_k^*()$
            $\delta \mathscr{U}_{\mathrm{p}}(\mathrm{q}, \mathrm{p})(x) = -n \mathbb{E}_{\mathrm{q}(\theta)}\left[\ell(\theta, x)\right] - \lambda \left\|x - \phi_k^*(x)\right\|^p$
          Update
            $\tilde{\boldsymbol{w}}_{k+1}^j = \boldsymbol{w}_k^j (1 - \eta_k \delta \mathscr{U}_{\mathrm{p}}(\mathrm{q}, \mathrm{p})(x_k^j))$
            $(\boldsymbol{w}_{k+1}^j)_{j \in [n]} = \text{reweighting}(\tilde{\boldsymbol{w}}_{k+1}^j)_{j \in [n]} \quad \mathrm{p}_{k+1} = \sum_{j=1}^n \boldsymbol{w}_{k+1}^j \delta_{x_{k+1}^j}$
    if $k = L + 1$ then
        Remove samples with smallest $n_r$ weights and assign equal weights of the rest of the
          samples.
    else
        $\mathrm{p}_{k+1} = \mathrm{p}_k$
return $\theta_T, x_T$;
```

C.2 Wasserstein-Fisher-Rao Gradient Descent Ascent

Algorithm 3: Wasserstein-Fisher-Rao Gradient Descent Ascent

```
Input: Initial particles $\theta_0^1,\cdots,\theta_0^m$, Data $x_0^1,\cdots,x_0^n$, Weights $\boldsymbol{w}_0^1,\cdots,\boldsymbol{w}_0^n$, $\mathrm{p}_0=\sum_{j=1}^n \boldsymbol{w}_0^j\delta_{x_0^i}$,
       Terminal time $T$, Step size $\eta_k$, Cut-off time $L$, Cut-off number $n_r$
for $k=0$ to $T-1$ do
    Update $\theta$
    Set $\mathbf{v}_k=0$;
    Use $n_k$ Monte Carlo samples to estimate;
        $\mathbb{E}_{\pi_{k,x}(x)}\left[s_\theta\left(x,\theta_k\right)\right] \approx \frac{1}{n_k}\sum_{i=1}^{n_k} s_\theta\left(x^i,\theta_k\right) \quad \text{s.t. } x^1,\cdots,x^{n_k} \overset{iid}{\sim} \mathrm{p}_k=\sum_{i=1}^{n}\boldsymbol{w}_k^i\delta_{x_k^i}.$
    Compute the corresponding estimator $\mathbf{v}_k$;
        $\theta_{k+1}^j=\theta_k^j+\eta_k n\mathbb{E}_{\pi_{k,x}(x)}\left[s_\theta\left(x,\theta_k^j\right)\right]+\sqrt{2\eta_k}\xi^j,$
    where $\xi^j\sim\mathcal{N}\left(0,I_d\right)$, i.i.d. $j\in[n]$;
        $\mathrm{q}_{k+1}(\theta)=\frac{1}{m}\sum_{j=1}^{m}\delta_{\theta_{k+1}^j}$
    Use $m_k$ Monte Carlo samples to estimate;
        $\mathbb{E}_{\mathrm{q}_k(\theta)}\left[s_x\left(x_k,\theta\right)\right] \approx \frac{1}{m_k}\sum_{j=1}^{m_k} s_x\left(x_k,\theta^j\right) \quad \text{s.t. } \theta^1,\cdots,\theta^{m_k} \overset{iid}{\sim} \frac{1}{m}\sum_{j=1}^{m}\delta_{\theta_k^j}.$
    Update x
    Run Sinkhorn algorithm to compute optimal coupling $\pi_k^*$, potential functions $\phi_k^*()$ and first
      variation:
        $\delta\mathscr{U}_\mathrm{p}(\mathrm{q},\mathrm{p})(x)=-n\mathbb{E}_{\mathrm{q}(\theta)}\left[\ell(\theta,x)\right]-\lambda\phi_k^*(x)$
        $\nabla\delta\mathscr{U}_\mathrm{p}(\mathrm{q}_k,\mathrm{p}_k)(x)=-n\mathbb{E}_{\mathrm{q}(\theta)}\left[\nabla_x\ell(\theta,x)\right]-\lambda(\nabla_x\|x-y\|^p)I\{(x,y)\in\mathrm{supp}(\pi_k^*)\}$
    Wasserstein step ;
        $x_{k+1}^j=x_{k\eta_k}^j-\frac{\eta_k}{\lambda}\nabla\delta\mathscr{U}_\mathrm{p}(\mathrm{q}_k,\mathrm{p}_k)(x_k^j)$
    if $k\leq L$ then
        Fisher-Rao step ;
            $\tilde{\boldsymbol{w}}_{k+1}^j=\boldsymbol{w}_k^j(1-\eta_k\delta\mathscr{U}_\mathrm{p}(\mathrm{q},\mathrm{p})(x_k^j))$
            $(\boldsymbol{w}_{k+1}^j)_{j\in[n]}=\mathrm{reweighting}(\tilde{\boldsymbol{w}}_{k+1}^j)_{j\in[n]} \quad \mathrm{p}_{k+1}=\sum_{j=1}^{n}\boldsymbol{w}_{k+1}^j\delta_{x_{k+1}^j}$
    if $k=L+1$ then
        Remove samples with smallest $n_r$ weights and assign equal weights of the rest of the
          samples.
    else
        $\mathrm{p}_{k+1}=\mathrm{p}_k$
return $\theta_T, x_T$;
```

C.3 Simulation

We study DRVB for Bayesian Logistic Regression model under different types of contamination including Huber contamination, added noise contamination, and a mixture of both. We generate i.i.d data from Logistic regression model and then generate synthetic contamination to features. Then run Algorithms 1, 2 and 3 and Langevin Monte Carlo on contaminated data.

Setup: Bayesian Logistic Regression model with 5 parameters $\theta \in \mathbb{R}^d$ and intercept $\alpha = 0$. $\theta \sim \mathcal{N}(0, I_5)$. Features $X_i \overset{iid}{\sim} \mathcal{N}(0, 900)$ for $i \in [100]$, $Y_i \sim \text{Bernoulli}(\text{logit}(X_i^\top \theta))$,i.i.d., where $\text{logit}(x) = \dfrac{1}{1+\exp(-x)}$.

We first apply LMC to the clean data to obtain samples from the clean posterior. This is the "gold standard" of our inference.

Huber contamination. In this simulation, we consider the standard Huber contamination setting with a contamination rate of 0.1. Specifically, we randomly replace 10% of the samples with i.i.d. Cauchy-distributed random variables at varying scales. We then apply Langevin Monte Carlo (LMC) as well as the proposed algorithms Algorithms 1, 2 and 3 to the contaminated data to obtain posterior samples. To assess robustness, we compute the Wasserstein-1 distance between the posterior samples obtained from the contaminated data and the ground-truth posterior, approximated by L

Figures 2a, 2b, 3a, 3b, 4a and 4b present the Wasserstein-1 distance between variational posterior samples and the ground truth for $p = 3, 4$ across different contamination levels and varying λ.

The WGDA (W2) result follow a typical pattern: it first decreases, then increases, and may even diverge when the contamination level is too high. In contrast, Fisher-Rao (FR) + LMC consistently performs better than W2 and converges reliably in all settings. The Fisher-Rao gradient flow efficiently detects and removes outliers, making it particularly effective in the Huber contamination setting. On the other hand, the Wasserstein gradient flow is less robust to extreme contamination. Combining Fisher-Rao and W2 preserves the advantages of Fisher-Rao in handling outliers while ensuring stability in inference.

Remark 1. *In most plots, W2 shows a first decreasing and then increasing trend as λ keeps increasing. Ideally, for Wasserstein Gradient Flow(step size is infinitesimally small, discretization error is negligible), the increase of λ initially decreases the W_1 distance. This occurs because bigger λ leads to a smaller ambiguity set which further induces a variational solution with smaller bias. However, once λ reaches the critical point where the ambiguity set shrinks to the set with smallest radius which still contains the clean empirical measurem, further increases in λ causes the ambiguity set to exclude the clean empirical measure, introducing significant bias.*

In practice, however, for WGDA (fixed step size for varying λ), the turning point occurs much earlier due to the compounding effect of discretization error, which grows as λ increases.

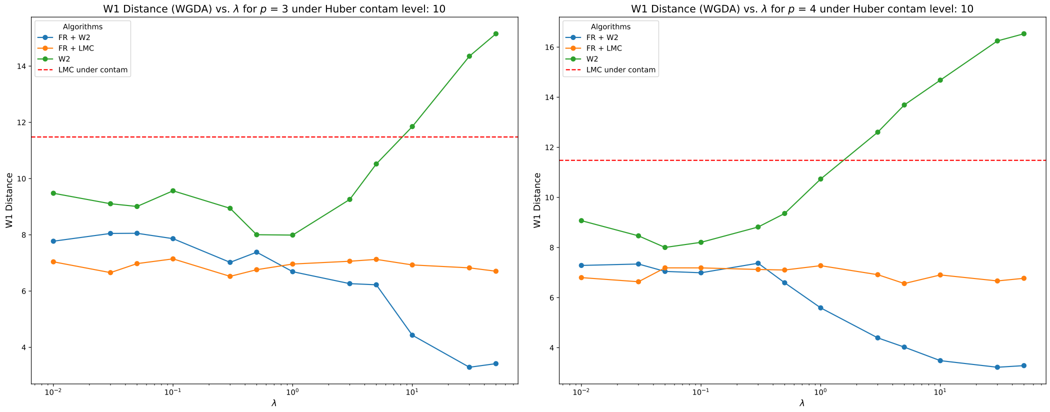
Added noise contamination. In this simulation, we consider the Added noise contamination setting. Contamination rate is 0.1. We randomly choose 10 percent samples and add i.i.d Laplace noise to them with different levels. Then we run Langevin Monte Carlo and Algorithms 1, 2 and 3 separately to the contaminated data to get corresponding posterior samples. Then we compute the Wasserstein-1 distance between the posterior samples we obtained by running algorithms on the contaminated samples and the ground truth(LMC samples from clean data). We study the cases that tuning parameter $p = 3, 4$ with varying λ.

Figures 5a, 5b, 6a, 6b, 7a and 7b display the W_1 distance between variational posterior samples and ground truth for $p = 3, 4$, varying λ and different levels.

Though Fisher-Rao is efficient to detect outliers, in the added-noise contamination setting, Fisher-Rao is less robust especially when noise level is relatively low. See Figures 5a and 5b, when noise level is 10, Fisher-Rao prone to take clean samples as contaminated samples leading to unstable results. While in this case, W2 is more robust as W2 mainly perturb the features which fits the contamination well. Fisher-Rao + W2 still preserves the robustness of W2 and performs nicely.

As the contamination level grows higher, the performance of Fisher-Rao gets a lot better since contaminated samples are easier to detect. And Fisher-Rao + W2 slightly outperforms W2.

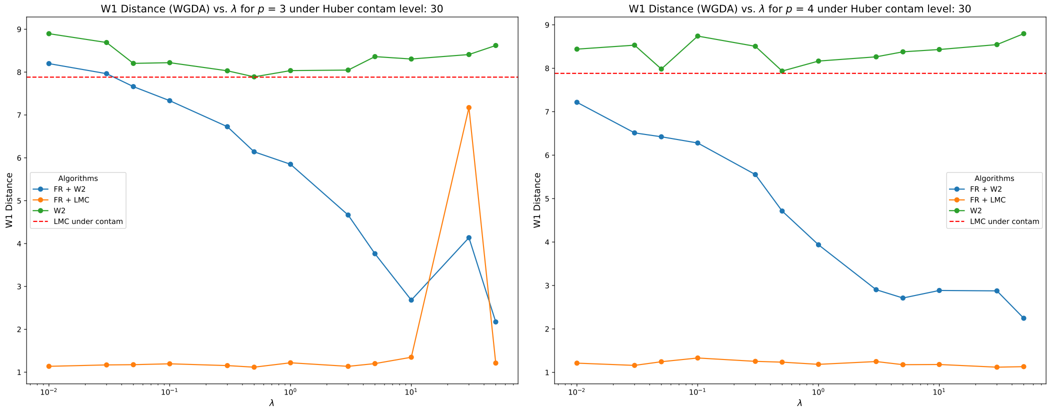
Mixed contamination. In this simulation, we consider a mixed contamination setting, a combination of Huber contamination and Added noise contamination. Contamination rate is still 0.1(0.05



(a) W_1 distance between LMC, WGDA, WFRGDA,FRGDA posterior samples and clean posterior samples for varying λ and $p = 3$ under Huber Contamination level 10

(b) W_1 distance between LMC, WGDA, WFRGDA, and FRGDA posterior samples and clean posterior samples for varying λ and $p = 4$ under Huber Contamination level 10

Figure 2: W_1 distance between posterior samples from LMC, WGDA, WFRGDA, and FGDA and clean posterior samples under different contamination levels.



(a) W_1 distance between LMC, WGDA, WFRGDA,FRGDA posterior samples and clean posterior samples for varying λ and $p = 3$ under Huber Contamination level 30

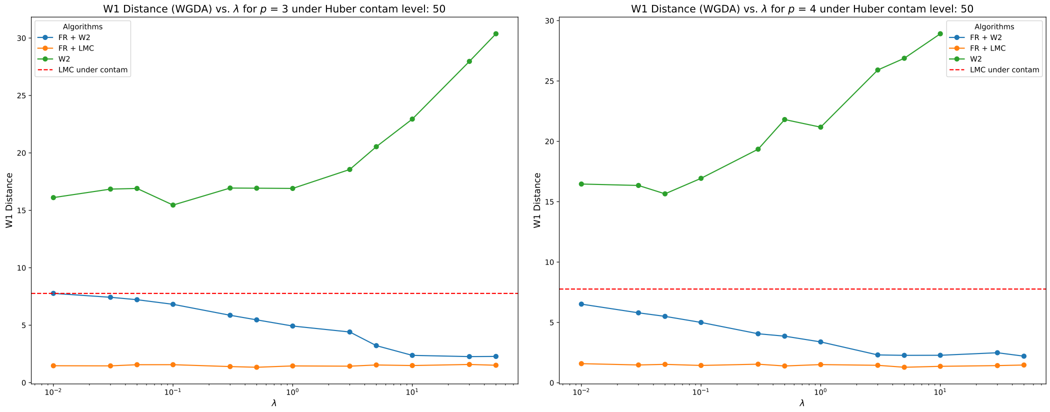
(b) W_1 distance between LMC, WGDA, WFRGDA, and FRGDA posterior samples and clean posterior samples for varying λ and $p = 4$ under Huber Contamination level 30

Figure 3: W_1 distance between posterior samples from LMC, WGDA, WFRGDA, and FGDA and clean posterior samples under different contamination levels.

for Huber and 0.05 for Added noise). We randomly choose 10 percent samples and substitute half of them by cauchy random variables and add i.i.d Laplace noise to the other half. Then we run 4 algorithms as before separately to the contaminated data to get posterior samples and compute the Wasserstein-1 distance between samples and the ground truth(LMC samples from clean data). We study the cases that tuning parameter $p = 3, 4$ with varying λ.

Figures 8a, 8b, 9a, 9b, 10a and 10b present the W_1 distance between variational posterior samples and ground truth for $p = 3, 4$, varying λ and different levels.

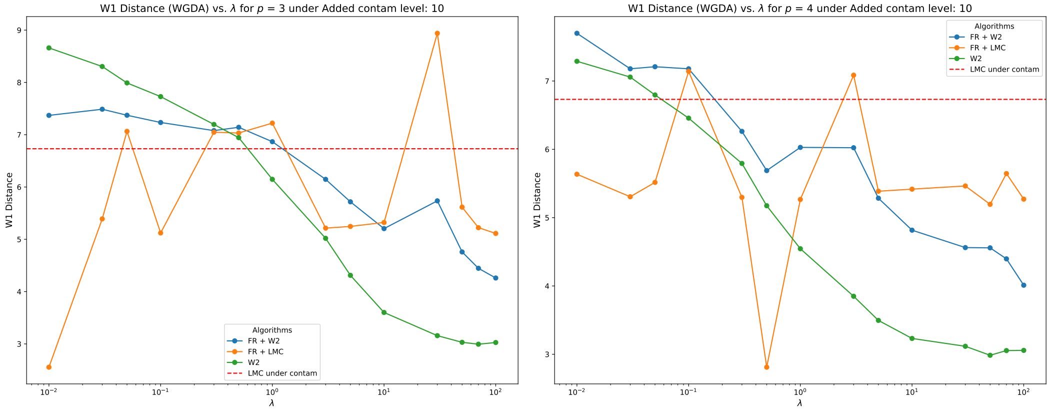
At level 10, Fisher-Rao is less robust than the other two algorithms due to the presence of weak added noise contamination. W2 and Fisher-Rao + W2 performs well in this case. And W2 with a larger λ performs best in all these three algorithms. As the contamination level increases to 30, Fisher-Rao performs the best and FR + W2 uniformly beats W2. W2 performs well with a relatively large λ. At the level 50, W2 can not handle the strong contamination any more. The performance is



(a) W_1 distance between LMC, WGDA, WFRGDA,FRGDA posterior samples and clean posterior samples for varying λ and $p = 3$ under Huber Contamination level 50

(b) W_1 distance between LMC, WGDA, WFRGDA, and FRGDA posterior samples and clean posterior samples for varying λ and $p = 4$ under Huber Contamination level 50

Figure 4: W_1 distance between posterior samples from LMC, WGDA, WFRGDA, and FGDA and clean posterior samples under contamination level 50.



(a) W_1 distance between LMC, WGDA, WFRGDA,FRGDA posterior samples and clean posterior samples for varying λ and $p = 3$ under Added Contamination level 10

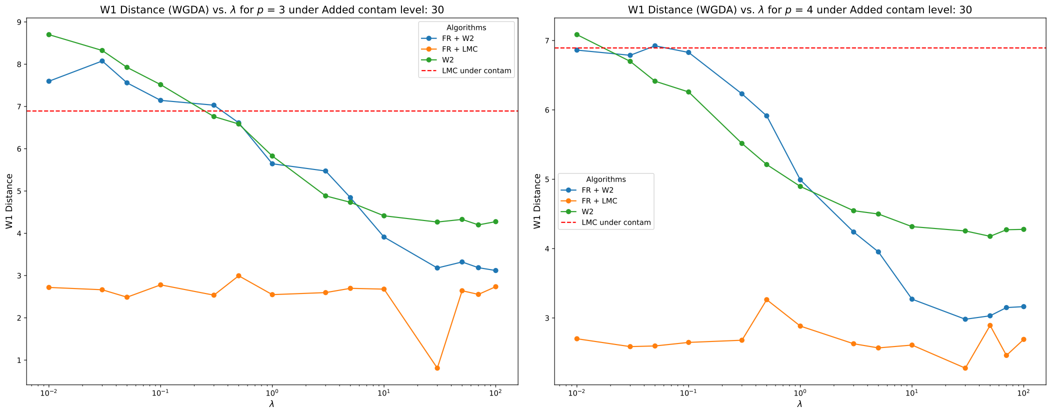
(b) W_1 distance between LMC, WGDA, WFRGDA, and FRGDA posterior samples and clean posterior samples for varying λ and $p = 4$ under Added Contamination level 10

Figure 5: W_1 distance between posterior samples from LMC, WGDA, WFRGDA, and FGDA and clean posterior samples under Added noise contamination level 10.

even worse than LMC. Fisher handles it well as it does in the single contamination case. And FR + W2 also performs robustly.

In summary, Fisher-Rao GDA is highly effective in handling Huber contamination and added noise contamination, particularly at high contamination levels. WGDA performs well in the presence of added noise contamination and remains effective under Huber contamination when the contamination level is low. When both types of contamination are present and contamination levels vary, the hybrid Fisher-Rao + Wasserstein (W2) approach demonstrates the highest robustness among all three algorithms.

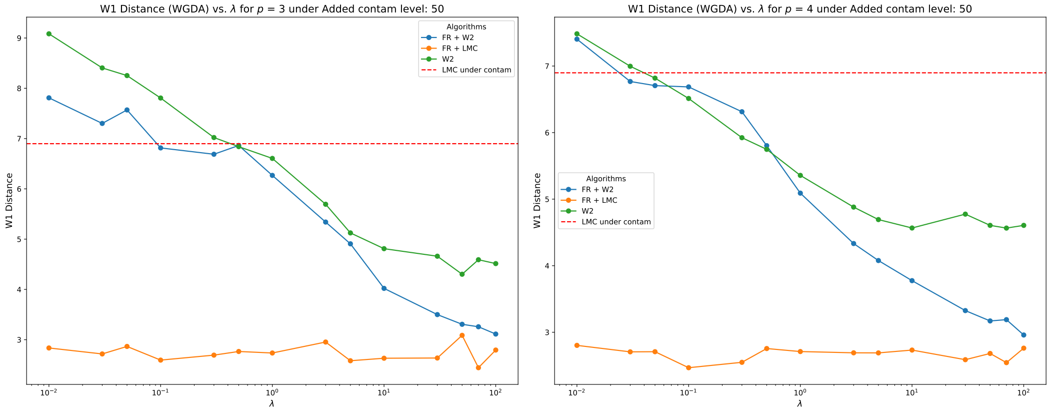
While there are scenarios where either W2 or Fisher-Rao individually may fail, Fisher-Rao + W2 consistently performs well and significantly outperforms LMC when properly tuned. This property can be interpreted as *double robustness*, meaning that the Fisher-Rao + W2 method remains effective as long as at least one of the two components is successful.



(a) W_1 distance between LMC, WGDA, WFRGDA,FRGDA posterior samples and clean posterior samples for varying λ and $p = 3$ under Added noise Contamination level 30

(b) W_1 distance between LMC, WGDA, WFRGDA, and FRGDA posterior samples and clean posterior samples for varying λ and $p = 4$ under Added Contamination level 30

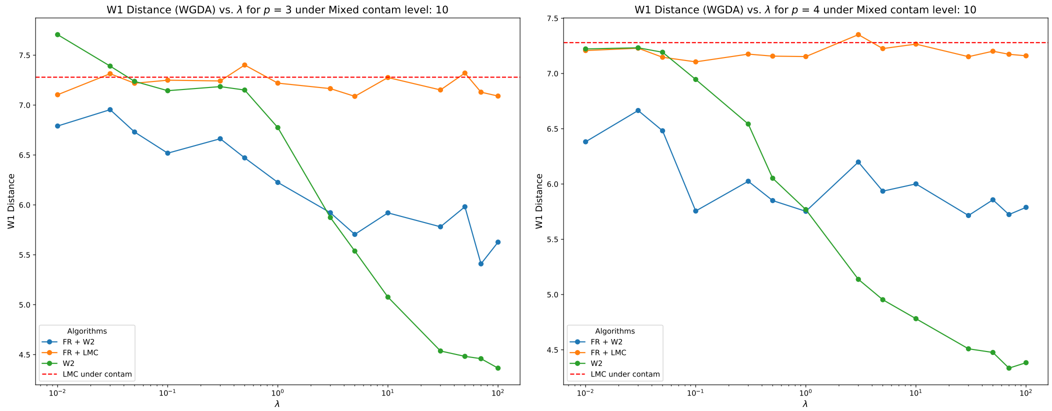
Figure 6: W_1 distance betweenposterior samples from LMC, WGDA, WFRGDA, and FGDA and clean posterior samples under Added contamination level 30.



(a) W_1 distance between LMC, WGDA, WFRGDA,FRGDA posterior samples and clean posterior samples for varying λ and $p = 3$ under Added Contamination level 50

(b) W_1 distance between LMC, WGDA, WFRGDA, and FRGDA posterior samples and clean posterior samples for varying λ and $p = 4$ under Added Contamination level 50

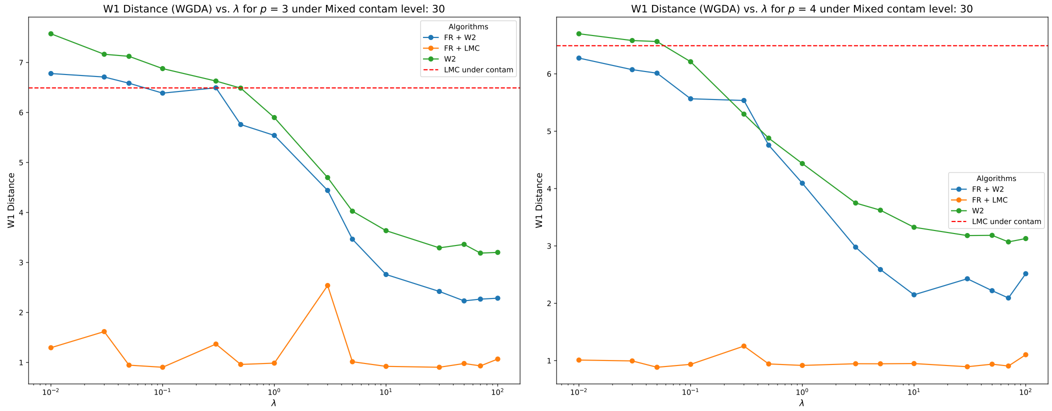
Figure 7: W_1 distance between posterior samples from LMC, WGDA, WFRGDA, and FGDA and clean posterior samples under contamination level 50.



(a) W_1 distance between LMC, WGDA, WFRGDA,FRGDA posterior samples and clean posterior samples for varying λ and $p = 3$ under Mixed Contamination level 10

(b) W_1 distance between LMC, WGDA, WFRGDA, and FRGDA posterior samples and clean posterior samples for varying λ and $p = 4$ under Huber Contamination level 10

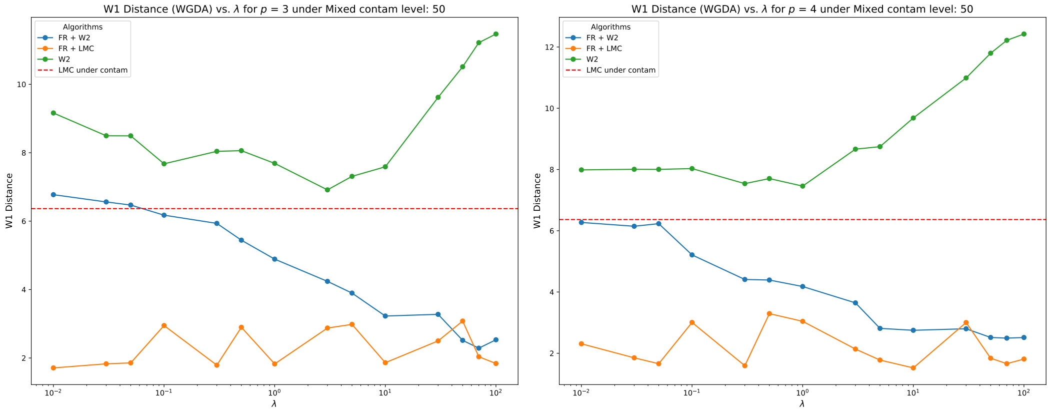
Figure 8: W_1 distance between posterior samples from LMC, WGDA, WFRGDA, and FGDA and clean posterior samples under Mixed contamination level 10.



(a) W_1 distance between LMC, WGDA, WFRGDA,FRGDA posterior samples and clean posterior samples for varying λ and $p = 3$ under Mixed Contamination level 30

(b) W_1 distance between LMC, WGDA, WFRGDA, and FRGDA posterior samples and clean posterior samples for varying λ and $p = 4$ under Huber Contamination level 30

Figure 9: W_1 distance betweenposterior samples from LMC, WGDA, WFRGDA, and FGDA and clean posterior samples under contamination level 30.



(a) W_1 distance between LMC, WGDA, WFRGDA,FRGDA posterior samples and clean posterior samples for varying λ and $p = 3$ under Mixed Contamination level 50

(b) W_1 distance between LMC, WGDA, WFRGDA, and FRGDA posterior samples and clean posterior samples for varying λ and $p = 4$ under Mixed Contamination level 50

Figure 10: W_1 distance between posterior samples from LMC, WGDA, WFRGDA, and FGDA and clean posterior samples under Mixed contamination level 50.