Conditional Generative Models are Provably Robust: Pointwise Guarantees for Bayesian Inverse Problems

Anonymous authors
Paper under double-blind review

Abstract

Conditional generative models became a very powerful tool to sample from Bayesian inverse problem posteriors. It is well-known in classical Bayesian literature that posterior measures are quite robust with respect to perturbations of both the prior measure and the negative log-likelihood, which includes perturbations of the observations. However, to the best of our knowledge, the robustness of conditional generative models with respect to perturbations of the observations has not been investigated yet. In this paper, we prove for the first time that appropriately learned conditional generative models provide robust results for single observations.

1 Introduction

Initiated by Szegedy et al. (2013), the vulnerability of deep neural networks (NNs) to adversarial attacks has been shown in many papers, see for instance (Carlini, 2020; Ortiz-Jimenez et al., 2020; Yuan et al., 2019). The vast majority of the literature is concerned with classification and related tasks like image segmentation. Here, gradient-based information is typically used in order to cross the discontinuous decision boundary of the classifier.

In this paper, we are interested in the solution of inverse problems by Bayesian methods. In (del Aguila Pla et al., 2023), it was shown that for Gaussian noise and a convex negative log prior, the maximum a-posteriori (MAP) estimation is stable with respect to the observations. This is no longer true for non-convex log priors, see our motivating example in the appendix. Concerning the robustness of end-to-end NN architectures, e.g., when learning the NN with a parameter constrained quadratic loss function between the true data and their NN reconstruction from corresponding observations, there exist ambivalent results in the literature. Antun et al. (2020) observed that deep learning for inverse problems comes with instabilities in the sense that „tiny, almost undetectable perturbations, both in the image and sampling domain, may result in severe artifacts in the reconstruction", while Genzel et al. (2023) attested in their comprehensive tests that „ deep-learning-based methods are at least as robust as TV minimization with respect to adversarial noise". The authors of (Gandikota et al., 2022) showed experimentally the sensitivity of NN to perturbations for the inverse problem of image deblurring.

We are not interested in end-to-end learning methods, but rather in learning the whole posterior distribution in Bayesian inverse problems by conditional generative NNs as proposed, e.g., in (Adler

& Öktem, 2018; Ardizzone et al., 2019; Batzolis et al., 2021; Hagemann et al., 2022). Addressing the posterior measure instead of end-to-end reconstructions has several advantages as illustrated in our example in the appendix. In particular, (samples of) the posterior can be used to provide additional information on the reconstructed data, for example on their uncertainty. Further, several robustness guarantees on the posterior were proved in the literature. One of the first results in the direction of stability with respect to the distance of observations was obtained in (Stuart, 2010) with respect to the Hellinger distance, see also (Dashti & Stuart, 2017). A very related question instead of perturbed observations concerns the approximations of forward maps, which was investigated in (Marzouk & Xiu, 2009). Furthermore, different prior measures were considered in (Hosseini, 2017; Hosseini & Nigam, 2017; Sullivan, 2017), where they also discuss the general case in Banach spaces. Two recent works (Latz, 2020; Sprungk, 2020) investigated the (Lipschitz) continuity of the posterior measures with respect to a multitude of metrics, where Latz (2020) focused on the well-posedness of the Bayesian inverse problem and Sprungk (2020) on the local Lipschitz continuity. Most recently, in (Garbuno-Inigo et al., 2023) the stability estimates have been generalized to integral probability metrics circumventing some Lipschitz conditions done in (Sprungk, 2020). Our paper is based on the findings in (Sprungk, 2020), but relates them with conditional generative NNs that aim to learn the posterior.

More precisely, in many machine learning papers, the following idea is pursued in order to solve inverse problems simultaneously for all observations y: Consider a family of generative models $G_\theta(y,\cdot)$ with parameters θ, which are supposed to map a latent distribution, like the standard Gaussian one, to the absolutely continuous posteriors $P_{X|Y=y}$, i.e., $G_\theta(y,\cdot)_\# P_Z \approx P_{X|Y=y}$. In order to learn such a conditional generative model, usually a loss of the form

$$L(\theta) \coloneqq \mathbb{E}_{y\sim P_Y}[D(P_{X|Y=y}, G_\theta(y,\cdot)_\# P_Z)]$$

is chosen with some „distance" D between measures like the Kullback-Leibler (KL) divergence $D = \mathrm{KL}$ used in (Ardizzone et al., 2019) or the Wasserstein-1 distance $D = W_1$ appearing, e.g., in the framework of (conditional) Wasserstein generative adversarial networks (GANs) (Adler & Öktem, 2018; Arjovsky et al., 2017; Liu et al., 2021). Also conditional diffusion models (Igashov et al., 2022; Song et al., 2021b; Tashiro et al., 2021) fit into this framework. Here De Bortoli (2022) showed that the standard score matching diffusion loss also optimizes the Wasserstein distance between the target and predicted distribution.

However, in practice we are usually interested in the reconstruction quality from a single or just a few measurements which are null sets with respect to P_Y. In this paper, we are interested in the important question, whether there exist any guarantees for the NN output to be close to the posterior for one specific measurement $\tilde{y}$. Our main result in Theorem 5 shows that for a NN learned such that the loss becomes small in the Wasserstein-1 distance, say $L(\theta) < \varepsilon$, the distance $W_1(P_{X|Y=\tilde{y}}, G_\theta(\tilde{y},\cdot)_\# P_Z)$ becomes also small for the single observation $\tilde{y}$. More precisely, we get the bound

$$W_1(P_{X|Y=\tilde{y}}, G_\theta(\tilde{y},\cdot)_\# P_Z) \leq C\varepsilon^{\frac{1}{n+1}},$$

where C is a constant and n is the dimension of the observations. To the best of our knowledge, this is the first estimate given in this direction.

We like to mention that in contrast to our paper, where we assume that samples are taken from the distribution for which the NN was learned, the authors of (Hong et al., 2022) observed that conditional normalizing flows are unstable when feeding them out-of-distribution observations. This

is not too surprising given some literature on the instability of (conditional) normalizing flows (Behrmann et al., 2021; Kirichenko et al., 2020).

Outline of the paper. The main theorem is shown in Section 2. For this we introduce several lemmata for the local Lipschitz continuity of posterior measures and conditional generative models with respect to the Wasserstein distance. In Section 3, we discuss the dependence of our derived bound on the training loss for different conditional generative models. In Appendix A, we illustrate by a simple example with a Gaussian mixture prior and Gaussian noise, why posterior distributions can be expected to be more stable than maximum a-posteriori (MAP) estimations and have more desirable properties than minimum mean squared error (MMSE) estimations.

2 Pointwise Robustness of Conditional Generative NNs

Let $X \in \mathbb{R}^m$ be a continuous random variable with law P_X determined by its density function p_X and $f\colon \mathbb{R}^m \to \mathbb{R}^n$ a measurable function. We consider a Bayesian inverse problem

$$Y = \text{noisy}(f(X)) \tag{1}$$

where "noisy" describes the underlying noise model. A typical choice is additive Gaussian noise, resulting in

$$Y = f(X) + \Xi, \quad \Xi \sim \mathcal{N}(0, \sigma^2 I_n).$$

Let $G_\theta = G\colon \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^m$ be a conditional generative model trained to approximate the posterior distribution $P_{X|Y=y}$ using the latent random variable $Z \in \mathbb{R}^d$. We will assume that all appearing measures are absolutely continuous and that the first moment of $G(y,\cdot)_\# P_Z$ is finite for all $y \in \mathbb{R}^n$. In particular, the posterior density is related via Bayes' theorem through the prior p_X and the likelihood $p_{Y|X=x}$ as

$$p_{X|Y=y} \propto p_{Y|X=x} p_X,$$

where $\propto$ means quality up to a multiplicative normalization constant. Further, we assume that the negative log-likelihood $-\log p_{Y|X=x}$ is bounded from below with respect to x, i.e., $\inf_x -\log p_{Y|X=x} > -\infty$. In particular, this includes mixtures of additive and multiplicative noise $Y = f(X) + \Xi_1 + \Xi_2 f(X)$, if X, Ξ_1 and Ξ_2 are independent, or log-Poisson noise commonly arising in computerized tomography.

We will use the Wasserstein-1 distance (Villani, 2009), which is a metric on the space of probability measures with finite first moment and is defined for measures μ and ν on the space $\mathbb{R}^m$ as

$$W_1(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^m \times \mathbb{R}^m} \|x - y\| d\pi(x,y),$$

where $\Pi(\mu,\nu)$ contains all measures on $\mathbb{R}^m \times \mathbb{R}^m$ with μ and ν as its marginals. The Wasserstein distance can be also rewritten by its dual formulation (Villani, 2009, Remark 6.5) as

$$W_1(\mu,\nu) = \max_{\text{Lip}(\varphi) \le 1} \int \varphi(x) \mathrm{d}(\mu - \nu)(x). \tag{2}$$

First, we show the Lipschitz continuity of our generating measures $G(y,\cdot)_\# P_Z$ with respect to y.

Lemma 1 (Local Lipschitz continuity of generator)**.** *For any parameterized family of generative models G with $\|\nabla_y G(y,z)\| \leq L_r$ for all $z \in \operatorname{supp}(P_Z)$ and all $y \in \mathbb{R}^n$ with $\|y\| \leq r$ for some $L_r > 0$ and some $r > 0$, it holds*

$$W_1(G(y_1,\cdot)_\# P_Z, G(y_2,\cdot)_\# P_Z) \leq L_r \|y_1 - y_2\|$$

for all $y_1, y_2 \in \mathbb{R}^n$ with $\|y_1\|, \|y_2\| \leq r$.

Proof. We use the mean value theorem which yields for every $z \in \operatorname{supp}(P_Z)$ and all $y_1, y_2 \in \mathbb{R}^n$ with $\|y_1\|, \|y_2\| \leq r$

$$\begin{aligned}
\|G(y_1,z) - G(y_2,z)\| &= \Big\| \int_0^1 \nabla_y G(y_1 + t(y_2 - y_1), z)(y_1 - y_2) \mathrm{d}t \Big\| \\
&\leq \int_0^1 \|\nabla_y G(y_1 + t(y_2 - y_1), z)\| \mathrm{d}t \|y_1 - y_2\| \\
&\leq L_r \|y_1 - y_2\|.
\end{aligned}$$

Next, we apply the dual formulation of the Wasserstein-1 distance to estimate

$$\begin{aligned}
W_1(G(y_1,\cdot)_\# P_Z, G(y_2,\cdot)_\# P_Z) &= \max_{\operatorname{Lip}(\varphi)\leq 1} \mathbb{E}_{z\sim P_Z}[\varphi(G(y_1,z)) - \varphi(G(y_2,z))] \\
&\leq \max_{\operatorname{Lip}(\varphi)\leq 1} \mathbb{E}_{z\sim P_Z}[|\varphi(G(y_1,z)) - \varphi(G(y_2,z))|] \\
&\leq \mathbb{E}_{z\sim P_Z}[\|G(y_1,z) - G(y_2,z)\|] \\
&\leq L_r \|y_1 - y_2\|.
\end{aligned}$$

□

Remark 2. *If P_Z is supported on a compact set, then the assumption in Lemma 1 is fulfilled for generators which are, e.g., continuously differentiable and then it follows from the extreme value theorem. Note that if we choose P_Z to be a Gaussian distribution, then it holds $\operatorname{supp}(P_Z) = \mathbb{R}^d$. Thus, for continuously differentiable generators it is not clear that this assumption is fulfilled, but at least the weaker assumption $\|\nabla_y G(y,z)\| \leq L_r$ for all $z \in \mathbb{R}^d$ with $\|z\| \leq \tilde{r}$ and all $y \in \mathbb{R}^n$ with $\|y\| \leq r$ holds true. In this case, we can show that Lemma 1 holds true up to an arbitrary small additive constant, see Appendix B for more details.*

By the following lemma, which is just (Sprungk, 2020, Corollary 19) for Euclidean spaces, the local Lipschitz continuity of the posterior distribution with respect to the Wasserstein-1 distance is guaranteed.

Lemma 3 (Local Lipschitz continuity of the posterior)**.** *Let the forward operator f and the likelihood $p_{Y|X=x}$ in (1) be measurable. Assume that there exists a function $M\colon [0,\infty) \times \mathbb{R} \to [0,\infty)$ which is monotone in the first component and non-decreasing in the second component such that for all $y_1, y_2 \in \mathbb{R}^n$ with $\|y_1\|, \|y_2\| \leq r$ for $r > 0$ and for all $x \in \mathbb{R}^m$ it holds*

$$|\log p_{Y|X=x}(y_2) - \log p_{Y|X=x}(y_1)| \leq M(r, \|x\|)\|y_1 - y_2\|. \tag{3}$$

Furthermore, assume that $M(r, \|\cdot\|) \in L^2_{P_X}(\mathbb{R}^m, \mathbb{R})$. Then, for any $r > 0$ there exists a constant $C_r < \infty$ such that for all $y_1, y_2 \in \mathbb{R}^n$ with $\|y_1\|, \|y_2\| \leq r$ we have

$$W_1(P_{X|Y=y_1}, P_{X|Y=y_2}) \leq C_r \|y_1 - y_2\|.$$

The Lipschitz constants of the family of generative models G^ε and the posterior distributions $P_{X|Y=y}$ can be related to each other under some convergence assumptions. Let the assumptions of Lemma 3 be fulfilled, assume further that

$$\lim_{\varepsilon\to 0} G^\varepsilon(y,\cdot)_\# P_Z = P_{X|Y=y}$$

with respect to the W_1-distance and consider observations $y_1, y_2 \in \mathbb{R}^n$ with $\|y_1,\|, \|y_2\| \leq r$. Then, by the triangle inequality it holds

$$\begin{aligned}\lim_{\varepsilon\to 0} W_1(G^\varepsilon(y_1,\cdot)_\# P_Z, G^\varepsilon(y_2,\cdot)_\# P_Z) &\leq \lim_{\varepsilon\to 0} W_1(G^\varepsilon(y_1,\cdot)_\# P_Z, P_{X|Y=y_1}) + W_1(P_{X|Y=y_1}, P_{X|Y=y_2}) \\ &\quad + W_1(P_{X|Y=y_2}, G^\varepsilon(y_2,\cdot)_\# P_Z) \\ &= W_1(P_{X|Y=y_1}, P_{X|Y=y_2}) \\ &\leq C_r \|y_1 - y_2\|.\end{aligned}$$

Hence, under the assumption of convergence, we expect the Lipschitz constant of our conditional generative models to behave similar to the one of the posterior distribution.

Remark 4. *The assumption* (3) *is for instance fulfilled for additive Gaussian noise* $\Xi \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$. *In this case*

$$-\log p_{Y|X=x}(y) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\|y - f(x)\|^2.$$

Hence $-\log p_{Y|X=x}(y)$ *is differentiable with respect to* y *and we get local Lipschitz continuity of the negative log-likelihood.*

Now we can prove our main theorem which ensures pointwise bounds on the distance between posterior and generated measure, if the training loss becomes small. In particular, the bound depends on the local Lipschitz constant of the conditional generator with respect to the observation, the local Lipschitz constant of the inverse problem, the training loss and the probability of the considered observation $\tilde{y}$. We want to highlight that the bound depends on the evidence $p_Y(\tilde{y})$ of an observation $\tilde{y}$ and indicates that we generally cannot expect a good pointwise estimate for out-of-distribution observations, i.e., $p_Y(y) \approx 0$. This is in agreement with the empirical results presented in (Hong et al., 2022).

Theorem 5. *Let the forward operator* f *and the likelihood* $p_{Y|X=x}$ *in* (1) *fulfill the assumptions of Lemma 3. Let* $\tilde{y} \in \mathbb{R}^n$ *be an observation with* $p_Y(\tilde{y}) = a > 0$. *Further, assume that* $y \mapsto p_Y(y)$ *is differentiable with* $\|\nabla p_Y(y)\| \leq K$ *for* $K > 0$ *and all* $y \in \mathbb{R}^n$. *Assume that we have trained a family of generative models* G *which fulfills* $\|\nabla_y G(y,z)\| \leq L_k$ *for all* $z \in \operatorname{supp}(P_Z)$ *and all* $y \in \mathbb{R}^n$ *with* $\|y\| \leq k$ *for some* $L_k > 0$ *and some* $k \geq \frac{a}{2K} + \|\tilde{y}\|$, *such that*

$$\mathbb{E}_{y\sim P_Y}[W_1(P_{X|Y=y}, G(y,\cdot)_\# P_Z)] \leq \varepsilon \tag{4}$$

for some $\varepsilon > 0$. *Then we have for* $\varepsilon \leq \left(\frac{a}{2K}\right)^{n+1} \frac{(L_{\|\tilde{y}\|+\frac{a}{2K}} + C_{\|\tilde{y}\|+\frac{a}{2K}})S_n a}{2n}$ *that*

$$W_1(P_{X|Y=\tilde{y}}, G(\tilde{y},\cdot)_\# P_Z) \leq (L_{\|\tilde{y}\|+\frac{a}{2K}} + C_{\|\tilde{y}\|+\frac{a}{2K}})^{1-\frac{1}{n+1}} \left(1 + \frac{1}{n}\right) \left(\frac{2n}{S_n a}\right)^{\frac{1}{n+1}} \varepsilon^{\frac{1}{n+1}}, \tag{5}$$

where $S_n \coloneqq \pi^{\frac{n}{2}}/\Gamma(\frac{n}{2}+1)$ *and* $C_\bullet$ *is the Lipschitz constant from Lemma 3. If* $\varepsilon \leq 1$, *it also holds*

$$W_1(P_{X|Y=\tilde{y}}, G(\tilde{y},\cdot)_\# P_Z) \leq (L_{\|\tilde{y}\|+\frac{a}{2K}} + C_{\|\tilde{y}\|+\frac{a}{2K}}) \frac{\varepsilon^{\frac{1}{n+1}} a}{2K} + \frac{2\varepsilon^{\frac{1}{n+1}}}{S_n (\frac{a}{2K})^n a}. \tag{6}$$

Proof. Let $0 < r \leq \frac{a}{2K}$. Then, for $y \in B_r(\tilde{y})$, there exists by the mean value theorem some $\xi \in \overline{y\tilde{y}}$ such that

$$|p_Y(y) - p_Y(\tilde{y})| \leq \|\nabla p_Y(\xi)\|\|y - \tilde{y}\| \leq Kr \leq \frac{a}{2}.$$

Consequently, each $y \in B_r(\tilde{y})$ has at least probability $p_Y(y) \geq \frac{a}{2}$. Moreover, by the volume of the n-dimensional ball it holds that

$$P_Y(B_r(\tilde{y})) = \int_{B_r(\tilde{y})} p_Y(y)\mathrm{d}y \geq \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} r^n \frac{a}{2} = S_n r^n \frac{a}{2}.$$

Now we claim that there exists $\widehat{y} \in B_r(\tilde{y})$ with

$$W_1(P_{X|Y=\widehat{y}}, G(\widehat{y}, \cdot)_{\#} P_Z) \leq \frac{2\varepsilon}{S_n r^n a}. \tag{7}$$

If this would not be the case, this would imply a contradiction to (4) by

$$\begin{aligned}
\mathbb{E}_{y\sim P_Y}[W_1(P_{X|Y=y}, G(y,\cdot)_{\#}P_Z)] &= \int_{\mathbb{R}^n} W_1(P_{X|Y=y}, G(y,\cdot)_{\#}P_Z)\mathrm{d}P_Y(y) \\
&\geq \int_{B_r(\tilde{y})} W_1(P_{X|Y=y}, G(y,\cdot)_{\#}P_Z)\mathrm{d}P_Y(y) \\
&> \int_{B_r(\tilde{y})} \frac{2\varepsilon}{S_n r^n a}\mathrm{d}P_Y(y) \\
&= P_Y(B_r(\tilde{y}))\frac{2\varepsilon}{S_n r^n a} \geq \varepsilon.
\end{aligned}$$

Next, we show the local Lipschitz continuity of $y \mapsto W_1(P_{X|Y=y}, G(y,\cdot)_{\#}P_Z)$ on $B_r(\tilde{y})$ by combining Lemma 1 and Lemma 3. Let $y_1, y_2 \in B_r(\tilde{y})$, so that $\|y_1\|, \|y_2\| \leq \|\tilde{y}\| + r$. Let $L_{\|\tilde{y}\|+r} > 0$ be the local Lipschitz constant from Lemma 1 and $C_{\|\tilde{y}\|+r}$ the local Lipschitz constant from Lemma 3. Using the triangle inequality and its reverse, we get

$$\begin{aligned}
&|W_1(P_{X|Y=y_1}, G(y_1,\cdot)_{\#}P_Z) - W_1(P_{X|Y=y_2}, G(y_2,\cdot)_{\#}P_Z)| \\
&\leq |W_1(P_{X|Y=y_1}, G(y_1,\cdot)_{\#}P_Z) - W_1(P_{X|Y=y_1}, G(y_2,\cdot)_{\#}P_Z)| \\
&\quad + |W_1(P_{X|Y=y_1}, G(y_2,\cdot)_{\#}P_Z) - W_1(P_{X|Y=y_2}, G(y_2,\cdot)_{\#}P_Z)| \\
&\leq W_1(G(y_1,\cdot)_{\#}P_Z, G(y_2,\cdot)_{\#}P_Z) + W_1(P_{X|Y=y_1}, P_{X|Y=y_2}) \\
&\leq (L_{\|\tilde{y}\|+r} + C_{\|\tilde{y}\|+r})\|y_1 - y_2\|.
\end{aligned} \tag{8}$$

Combination of the results in (7) and (8) yields the estimate

$$\begin{aligned}
W_1(P_{X|Y=\tilde{y}}, G(\tilde{y},\cdot)_{\#}P_Z) &\leq |W_1(P_{X|Y=\tilde{y}}, G(\tilde{y},\cdot)_{\#}P_Z) - W_1(P_{X|Y=\widehat{y}}, G(\widehat{y},\cdot)_{\#}P_Z)| \\
&\quad + |W_1(P_{X|Y=\widehat{y}}, G(\widehat{y},\cdot)_{\#}P_Z)| \\
&\leq (L_{\|\tilde{y}\|+r} + C_{\|\tilde{y}\|+r})r + \frac{2\varepsilon}{S_n r^n a} \\
&\leq (L_{\|\tilde{y}\|+\frac{a}{2K}} + C_{\|\tilde{y}\|+\frac{a}{2K}})r + \frac{2\varepsilon}{S_n r^n a}.
\end{aligned}$$

The radius r, for which the right-hand side becomes minimal, is given by

$$r = \Big(\frac{2n\varepsilon}{(L_{\|\tilde{y}\|+\frac{a}{2K}} + C_{\|\tilde{y}\|+\frac{a}{2K}})S_n a}\Big)^{\frac{1}{n+1}}.$$

Plugging this in, we get (5). However, we need that $r \leq \frac{a}{2K}$ which implies

$$\varepsilon \leq \Big(\frac{a}{2K}\Big)^{n+1} \frac{(L_{\|\tilde{y}\|+\frac{a}{2K}} + C_{\|\tilde{y}\|+\frac{a}{2K}})S_n a}{2n}.$$

On the other hand, if $\varepsilon \leq 1$, we can choose $r = \varepsilon^{\frac{1}{n+1}} \frac{a}{2K} \leq \frac{a}{2K}$ which results in (6) and has the same asymptotic rate. □

Remark 6. *We can get rid of the dimension scaling $\varepsilon^{\frac{1}{n+1}}$ by choosing the radius as $r = \frac{a}{2K}$, which yields*

$$W_1(P_{X|Y=\tilde{y}}, G(\tilde{y}, \cdot)_\# P_Z) \leq (L_{\|\tilde{y}\|+\frac{a}{2K}} + C_{\|\tilde{y}\|+\frac{a}{2K}})\frac{a}{2K} + \frac{2\varepsilon}{S_n(\frac{a}{2K})^n a}.$$

This comes at the disadvantage that the first term is constant with respect to ε.

The following corollary provides a characterization of a perfect generative model. If the expectation (4) goes to zero, then for all $y \in \mathbb{R}^n$ with $p_Y(y) > 0$ the posteriors $P_{X|Y=y}$ get predicted correctly.

Corollary 7. *Let the assumptions of Lemma 1 and Lemma 3 hold true and assume a global Lipschitz constant in Lemma 1. Let p_Y be differentiable with $\|\nabla p_Y(y)\| \leq K$ for some $K > 0$ and all $y \in \mathbb{R}^n$. Consider a family of generative networks $(G^\varepsilon)_{\varepsilon>0}$ fulfilling*

$$\mathbb{E}_{y\sim P_Y}[W_1(P_{X|Y=y}, G^\varepsilon(y, \cdot)_\# P_Z)] \leq \varepsilon$$

and assume that the Lipschitz constants L^ε of G^ε from Lemma 1 are bounded by some $L < \infty$. Then for all observations $y \in \mathbb{R}^n$ with $p_Y(y) > 0$ it holds

$$W_1(P_{X|Y=y}, G^\varepsilon(y, \cdot)_\# P_Z) \to 0 \quad \text{as } \varepsilon \to 0.$$

Proof. We can assume that $\varepsilon \leq 1$, then the statement follows immediately from Theorem 5. □

3 Conditional Generative Models

In this section, we discuss whether the main assumption, namely that the averaged Wasserstein distance $\mathbb{E}_{y\sim P_Y}[W_1(P_{X|Y=y}, G(y, \cdot)_\# P_Z)]$ in (4) becomes small, is reasonable for different conditional generative models. Therefore we need to relate the typical choices of training loss with the Wasserstein distance.

3.1 Conditional Normalizing Flows

Conditional normalizing flows (Altekrüger & Hertrich, 2023; Andrle et al., 2021; Ardizzone et al., 2019; Winkler et al., 2019) are a family of normalizing flows parameterized by a condition, which in our case is the observation y. The aim is to learn a network $\mathcal{T}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ such that $\mathcal{T}(y, \cdot)$ is a

diffeomorphism and $\mathcal{T}(y,\cdot)_{\#}P_Z \approx P_{X|Y=y}$ for all $y \in \mathbb{R}^n$, where $\approx$ means that two distributions are similar in some proper distance or divergence. This can be done via minimizing the expectation on Y of the *forward* KL divergence $\mathbb{E}_{y\sim P_Y}[\mathrm{KL}(P_{X|Y=y}, \mathcal{T}(y,\cdot)_{\#}P_Z)]$, which is equal, up to a constant, to

$$\mathbb{E}_{x\sim P_X, y\sim P_Y}[-\log p_Z(\mathcal{T}^{-1}(y,x)) - \log(|\det D\mathcal{T}^{-1}(y,x)|)],$$

where the inverse is meant with respect to the second component, see (Hagemann et al., 2022) for more details. Training a network using the forward KL has many desirable properties like a mode-covering behaviour of $\mathcal{T}(y,\cdot)_{\#}P_Z$. Now conditional normalizing flows are trained using the KL divergence, while the theoretical bound in Section 2 relies on the metric properties of the Wasserstein-1 distance. Thus we need to show that we can ensure a small ε in (4) when training the conditional normalizing flow as proposed. Following (Gibbs & Su, 2002, Theorem 4), we can bound the Wasserstein distance by the total variation distance, which in turn is bounded by KL via Pinsker's inequality (Pinsker, 1963), i.e.,

$$\begin{aligned}\mathbb{E}_{y\sim P_y}[W_1((P_{X|Y=y}, \mathcal{T}(y,\cdot)_{\#}P_Z)^2] &\leq C\ \mathbb{E}_{y\sim P_Y}[\mathrm{TV}((P_{X|Y=y} - \mathcal{T}(y,\cdot)_{\#}P_Z)^2] \\ &\leq \frac{C}{\sqrt{2}}\mathbb{E}_{y\sim P_Y}[\mathrm{KL}((P_{X|Y=y}, \mathcal{T}(y,\cdot)_{\#}P_Z)],\end{aligned}$$

where C is a constant depending on the support of the probability measures. However, by definition $\mathrm{supp}(\mathcal{T}(y,\cdot)_{\#}P_Z) = \mathbb{R}^m$. By (Altekrüger et al., 2022, Lemma 4) the density $p_{\mathcal{T}(y,\cdot)_{\#}P_Z}$ decays exponentially. Therefore, we expect in practice that the Wasserstein distance becomes small if the KL vanishes even though (Gibbs & Su, 2002, Theorem 4) is not applicable.

3.2 Conditional Wasserstein GANs

In Wasserstein GANs (Arjovsky et al., 2017), a generative adversarial network approach is taken in order to sample from a target distribution. For this, the dual formulation (2) is used in order to calculate the Wasserstein distance between measures P_X and P_Y. Then the 1-Lipschitz function is reinterpreted as a discriminator in the GAN framework (Goodfellow et al., 2014). If the corresponding minimizer in the space of 1-Lipschitz functions can be found, then optimizing the adversarial Wasserstein GAN loss directly optimizes the Wasserstein distance. The classical Wasserstein GAN loss for a target measure μ and a generator $G\colon \mathbb{R}^d \to \mathbb{R}^m$ is given by

$$\min_{\theta} \max_{\mathrm{Lip}(\varphi)\leq 1} \mathbb{E}_{x\sim P_X, z\sim P_Z}[\varphi(x) - f(G(z))],$$

where $d \in \mathbb{N}$ is the dimension of the latent space.

The Wasserstein GAN framework can be extended to conditional Wasserstein GANs (Adler & Öktem, 2018; Liu et al., 2021) for solving inverse problems. For this, we aim to train generators $G\colon \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^m$ and average with respect to the observations

$$L(\theta) = \mathbb{E}_{y\sim P_y}\big[\max_{\mathrm{Lip}(\varphi_y)\leq 1} \mathbb{E}_{x\sim P_{X|Y=y}, z\sim P_Z}[\varphi_y(x) - \varphi_y(G(y,z))]\big].$$

Hence minimizing this loss (or a variant of it) directly enforces a small ε in assumption (4).

3.3 Conditional Diffusion Models

In diffusion models, a forward SDE, which maps a data distribution to an approximate Gaussian distribution is considered (Song et al., 2021a;b). Then the theory of reverse SDEs (Anderson, 1982) allows to sample from the data distribution by learning the score $\nabla \log p_t(x)$, where $p_t(x)$ is the path density of the forward SDE. The forward SDE usually reads

$$dX_t = -\alpha X_t \mathrm{d}t + \sqrt{2\alpha}\mathrm{d}W_t,$$

while the reverse SDE is given by

$$dY_t = -\alpha Y_t \mathrm{d}t - 2\ \nabla \log p_t(x)\mathrm{d}t + \sqrt{2\alpha}\mathrm{d}\tilde{W}_t,$$

where $\alpha \in \mathbb{R}$ describes the schedule of the SDE. However, the path density $p_t(x)$ is usually intractable, so that the score $\nabla \log p_t(x)$ is learned with a NN $s_\theta \colon [0,T] \times \mathbb{R}^m \to \mathbb{R}^m$ such that $s_\theta(t,x) \approx \nabla \log p_t(x)$ for all $t \in [0,T]$ and $x \in \mathbb{R}^m$. This can be ensured using the so-called score matching loss (Song et al., 2021b) defined by

$$\min_\theta \mathbb{E}_{t\sim U([0,T]), x\sim P_{X_t}} \left[\|s_\theta(t,x) - \nabla \log p_t(x)\|^2\right].$$

In order to solve inverse problems, we can consider a conditional reverse SDE

$$dY_t = -\alpha Y_t \mathrm{d}t - 2\ \nabla \log p_t(x|y)\mathrm{d}t + \sqrt{2\alpha}\mathrm{d}\tilde{W}_t,$$

where $p_t(x|y)$ is the conditional path density given an observation $y \in \mathbb{R}^n$. Consequently, we consider conditional diffusion models, where a NN $s_\theta \colon \mathbb{R}^n \times [0,T] \times \mathbb{R}^m \to \mathbb{R}^m$ is learned to approximate $s_\theta(y,t,x) \approx \nabla \log p_t(x|y)$ for all $t \in [0,T]$, $x \in \mathbb{R}^m$ and all observations $y \in \mathbb{R}^m$. Then the score matching loss for conditional diffusion models is given by (Batzolis et al., 2021, Theorem 1) as

$$L(\theta) = \mathbb{E}_{y\sim P_Y}\left[\mathbb{E}_{t\sim U([0,T]), x\sim P_{X_t|Y=y}}[\|s_\theta(y,t,x) - \nabla \log p_t(x|y)\|^2]. \right. \tag{9}$$

Denote by $\tilde{Y}$ the solution to the approximated SDE starting at $\tilde{Y}_0 \approx P_Z$ and $\tilde{Y}^y$ the solution of the approximated SDE conditioned on an observation $y \in \mathbb{R}^n$. Then we can use the bound derived in (Pidstrigach et al., 2023, Theorem 2) which gives

$$\mathbb{E}_{y\sim P_Y}[W_2(P_{X|Y=y}, P_{\tilde{Y}^y_T})] \leq \mathbb{E}_{y\sim P_Y}\left[C\ W_2\big(P_{X^y_T}, \mathcal{N}(0,\mathrm{Id})\big)\right] + TL(\theta),$$

where C is a constant depending on the length of the interval T and the Lipschitz constant of the conditional score $\nabla \log p_t(x|y)$. Finally, Hölders inequality yields for the Wasserstein-1 distance

$$\mathbb{E}_{y\sim P_Y}[W_1(P_{X|Y=y}, P_{\tilde{Y}^y_T})] \leq \mathbb{E}_{y\sim P_Y}\left[C\ W_2\big(P_{X^y_T}, \mathcal{N}(0,\mathrm{Id})\big)\right] + TL(\theta).$$

Hence, when training the conditional diffusion model by minimizing (9) we also ensure that (4) becomes small. For more in depth discussion with less restrictive assumptions on the score, see also (De Bortoli, 2022).

3.4 Conditional Variational Autoencoder

Variational Autoencoder (VAE) (Kingma & Welling, 2013) aim to approximate a distribution P_X by learning a stochastic encoder $E_\phi \colon \mathbb{R}^m \to \mathbb{R}^d \times \mathbb{R}^{d,d}$ determining parameters of the normal distribution $(\mu_\phi(x), \Sigma_\phi(x))$ for x sampled from P_X and pushing P_X to a latent distribution P_Z with density p_Z of dimension $d \in \mathbb{N}$. In the reverse direction, a stochastic decoder $D_\theta \colon \mathbb{R}^d \to \mathbb{R}^m \times \mathbb{R}^{m,m}$ determines parameters of the normal distribution $(\mu_\theta(z), \Sigma_\theta(z))$ for $z \in \mathbb{R}^d$ and pushes P_Z back to P_X. By definition, the densities of E_ϕ and D_θ are given by $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x))$ and $p_\theta(x|z) = \mathcal{N}(x; \mu_\theta(z), \Sigma_\theta(z))$, respectively. These networks are trained by minimizing the so-called evidence lower bound (ELBO)

$$\mathrm{ELBO}(\theta, \phi) = -\mathbb{E}_{x \sim P_X}\left[\mathbb{E}_{z \sim q_\phi(\cdot|x)}[\log(p_\theta(x|z)p_Z(z)) - \log(q_\phi(z|x)))]\right].$$

By (Hagemann et al., 2023, Theorem 4.1), the loss $L(\theta, \phi)$ is related to KL by

$$\mathrm{KL}(P_X, D_{\theta\#}P_Z) \leq \mathrm{ELBO}(\theta, \phi).$$

We can solve inverse problems by extending VAEs to conditional VAEs (Lim et al., 2018; Sohn et al., 2015) and aim to approximate the posterior distribution $P_{X|Y=y}$ for a given observation $y \in \mathbb{R}^n$. The conditional stochastic encoder $E_\phi \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^d \times \mathbb{R}^{d,d}$ and conditional stochastic decoder $D_\theta \colon \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^m \times \mathbb{R}^{m,m}$ are trained by

$$L(\theta, \phi) = \mathbb{E}_{y \sim P_Y}\left[-\mathbb{E}_{x \sim P_{X|Y=y}}\left[\mathbb{E}_{z \sim q_\phi(\cdot|y,x)}[\log(p_\theta(x|y,z)p_Z(z)) - \log(q_\phi(z|y,x)))]\right]\right].$$

By the same argument as above, the KL can be bounded by

$$\mathbb{E}_{y \sim P_Y}[\mathrm{KL}(P_{X|Y=y}, D_\theta(y,\cdot)_\# P_Z)] \leq L(\theta, \phi)$$

and, using similar arguments as in Section 3.1, we get the estimate

$$\mathbb{E}_{y \sim P_Y}[W_1(P_{X|Y=y}, D_\theta(y,\cdot)_\# P_Z)^2] \leq \frac{C}{\sqrt{2}} L(\theta, \phi).$$

4 Conclusion

We showed a pointwise stability guarantee of the Wasserstein distance between the posterior $P_{X|Y=y}$ of a Bayesian inverse problem and the learned distribution $G(y,\cdot)_\# P_Z$ of a conditional generative model G under certain assumptions. In particular, the pointwise bound depends on the Lipschitz constant of the conditional generator with respect to the observation, the Lipschitz constant of the inverse problem, the training loss with respect to the Wasserstein distance and the probability of the considered observation.

The required training accuracy of the bound depends on the Wasserstein-1 distance between the target distribution and the learned distribution. However, some conditional networks as the conditional normalizing flow are not trained to minimize the Wasserstein-1 distance. Consequently, a direct dependence of the bound on the training accuracy with respect to the KL divergence would be helpful. Under very strong assumptions, the continuity in Lemma 1 has been shown for KL in (Baptista et al., 2023). This could be used to derive a similar statement.

Furthermore, our bound is a worst case bound and is not always practical if the constants are large. It would be interesting to check whether tightness of the bound can be shown for some examples.

References

Jonas Adler and Ozan Öktem. Deep Bayesian inversion. *arXiv preprint arXiv:1811.05910*, 2018.

Fabian Altekrüger and Johannes Hertrich. WPPNets and WPPFlows: The power of Wasserstein patch priors for superresolution. *SIAM Journal on Imaging Sciences*, 2023.

Fabian Altekrüger, Alexander Denker, Paul Hagemann, Johannes Hertrich, Peter Maass, and Gabriele Steidl. PatchNR: learning from very few images by patch normalizing flow regularization. *arXiv preprint arXiv:2205.12021*, 2022.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Anna Andrle, Nando Farchmin, Paul Hagemann, Sebastian Heidenreich, Victor Soltwisch, and Gabriele Steidl. Invertible neural networks versus MCMC for posterior reconstruction in grazing incidence x-ray fluorescence. In Abderrahim Elmoataz, Jalal Fadili, Yvain Quéau, Julien Rabin, and Loïc Simon (eds.), *Scale Space and Variational Methods in Computer Vision*, pp. 528–539. Springer International Publishing, 2021.

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.

Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*, 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223. PMLR, 2017.

Ricardo Baptista, Bamdad Hosseini, Nikola B. Kovachki, Youssef M. Marzouk, and Amir Sagiv. An approximation theory framework for measure-transport sampling algorithms. *arXiv preprint arXiv:2302.13965*, 2023.

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger Grosse, and Joern-Henrik Jacobsen. Understanding and mitigating exploding inverses in invertible neural networks. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1792–1800. PMLR, 2021.

N. Carlini. A complete list of all (arxiv) adversarial example papers. 2020. URL `https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html?`

M. Dashti and A. M. Stuart. The a Bayesian approach to inverse problems. In *Handbook of Uncertainty Quantification*, pp. 311–428. Springer, 2017.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

P. del Aguila Pla, S. Neumayer, and M. Unser. Stability of image-reconstruction algorithms. *IEEE Transactions on Computational Imaging*, 2023.

Kanchana Vaishnavi Gandikota, Paramanand Chandramouli, and Michael Moeller. On adversarial robustness of deep image deblurring. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3161–3165, 2022.

Alfredo Garbuno-Inigo, Tapio Helin, Franca Hoffmann, and Bamdad Hosseini. Bayesian posterior perturbation analysis with integral probability metrics. *arXiv preprint arXiv:2303.01512*, 2023.

Martin Genzel, Jan Macdonald, and Maximilian März. Solving inverse problems with deep neural networks – robustness included? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1119–1134, 2023.

Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3):419–435, 2002.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3, 2014.

Dario Grana, Torstein Fjeldstad, and Henning Omre. Bayesian Gaussian mixture linear inversion for geophysical inverse problems. *Mathematical Geosciences*, 49(4):493–515, 2017.

P. Hagemann, J. Hertrich, and G. Steidl. *Generalized Normalizing Flows via Markov Chains*. Elements in Non-local Data Interactions: Foundations and Applications. Cambridge University Press, 2023.

Paul Hagemann, Johannes Hertrich, and Gabriele Steidl. Stochastic normalizing flows for inverse problems: A Markov chains viewpoint. *SIAM/ASA Journal on Uncertainty Quantification*, 10 (3):1162–1190, 2022.

Seongmin Hong, Inbum Park, and Se Young Chun. On the robustness of normalizing flows for inverse problems in imaging. *arXiv preprint arXiv:2212.04319*, 2022.

William C. Horrace. Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis*, 94(1):209–221, 2005.

B. Hosseini. Well-posed bayesian inverse problems with infinitely divisible and heavy-tailed prior measures. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):1024–1060, 2017.

B. Hosseini and N. Nigam. Well-posed Bayesian inverse problems: priors with exponential tails. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):436–465, 2017.

Ilia Igashov, Hannes Stärk, Clément Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design. *arXiv preprint arXiv:2210.05274*, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

Jonas Latz. On the well-posedness of Bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):451–482, 2020.

Jaechang Lim, Seongok Ryu, Jin Woo Kim, and Woo Youn Kim. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of cheminformatics*, 10(1):1–9, 2018.

Shiao Liu, Xingyu Zhou, Yuling Jiao, and Jian Huang. Wasserstein generative learning of conditional distribution. *arXiv preprint arXiv:2112.10039*, 2021.

Y. Marzouk and D. Xiu. A stochastic collocation approach to Bayesian inference in inverse problems. *Communications in Computational Physics*, 6(4):826–847, 2009.

G. Ortiz-Jimenez, A. Modas anmd S.-M. Moosavi-Dezfooli, and P. Frossard. Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness. *arXiv preprint arXiv:2010.09624*, 2020.

Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-dimensional diffusion models for function spaces. *arXiv preprint arXiv:2302.10130*, 2023.

MS Pinsker. Information and information stability of random quantities and processes. Technical report, Foreign Technology DIV Wright-Patterson AFB, Ohio, 1963.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, Montreal, Quebec, Canada*, pp. 3483–3491, 2015.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Björn Sprungk. On the local Lipschitz stability of Bayesian inverse problems. *Inverse Problems*, 36 (5), 2020.

A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

T. J. Sullivan. Well-posed Bayesian inverse problems and heavy-tailed stable quasi-banach space priors. *Inverse Problems in Imaging*, 11(5):857–874, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

G. M. Tallis. Elliptical and radial truncation in normal populations. *Annals of Mathematical Statistics*, 34:940–944, 1963.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Cédric Villani. *Optimal Transport: Old And New*, volume 338. Springer, 2009.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824, 2019.

A Example on the Robustness of the MAP and Posterior

We like to provide an example that illustrates the stability of the posterior distribution in contrast to the MAP estimator and highlights the role of the MMSE estimator.

By the following lemma, see, e.g., (Grana et al., 2017; Hagemann et al., 2023), the posterior of a Gaussian mixture model given observations from a linear forward operator corrupted by white Gaussian noise can be computed analytically.

Lemma 8. *Let $X \sim \sum_{k=1}^{K} w_k \mathcal{N}(m_k, \Sigma_k) \in \mathbb{R}^m$ be a Gaussian mixture random variable. Suppose that*

$$Y = AX + \Xi,$$

where $A : \mathbb{R}^m \to \mathbb{R}^n$ is a linear operator and $\Xi \sim N(0, \sigma^2 I_n)$. Then the posterior is also a Gaussian mixture

$$P_{X|Y=y} \propto \sum_{k=1}^{K} \tilde{w}_k \mathcal{N}(\cdot|\tilde{m}_k, \tilde{\Sigma}_k)$$

with

$$\tilde{\Sigma}_k \coloneqq (\tfrac{1}{\sigma^2} A^{\mathrm{T}} A + \Sigma_k^{-1})^{-1}, \qquad \tilde{m}_k \coloneqq \tilde{\Sigma}_k (\tfrac{1}{b^2} A^{\mathrm{T}} y + \Sigma_k^{-1} \mu_k)$$

and

$$\tilde{w}_k \coloneqq w_k \exp\left(\frac{1}{2}(\tilde{m}_k \tilde{\Sigma}_k^{-1} \tilde{m}_k - m_k \Sigma_k^{-1} m_k)\right).$$

Now, for some small $\varepsilon > 0$ we consider the random variable $X \in \mathbb{R}$ with simple prior distribution

$$P_X = \frac{1}{2}\mathcal{N}(-1, \varepsilon^2) + \frac{1}{2}\mathcal{N}(1, \varepsilon^2)$$

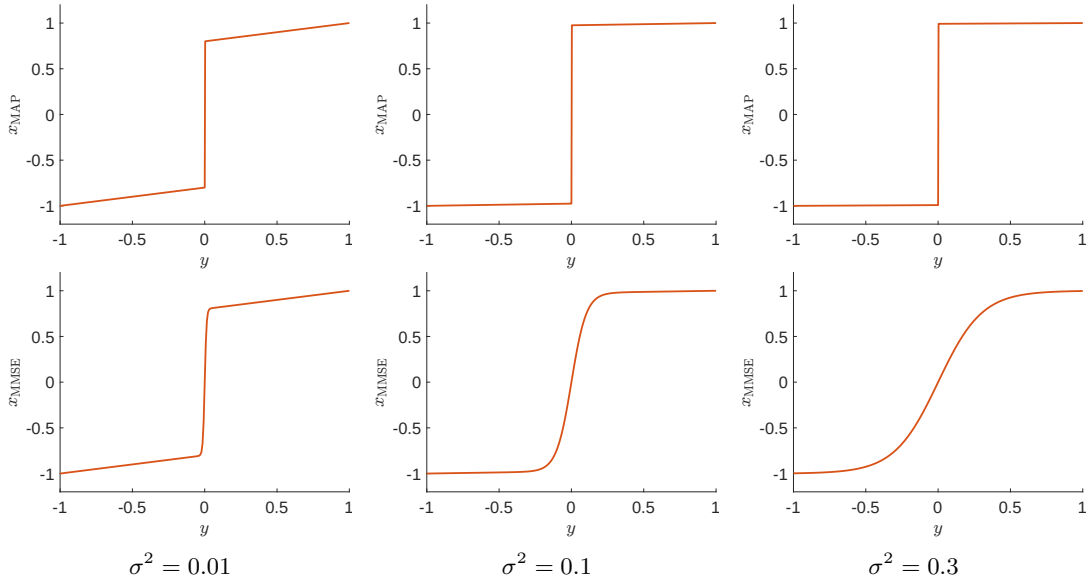


Figure 1: The MAP estimator (top) and the MMSE estimator (bottom) with respect to the observation y for $\varepsilon^2 = 0.05^2$ and different noise levels σ^2.

and observations from $Y = X + \Xi$ with noise $\Xi \sim \mathcal{N}(0, \sigma^2)$. The MAP estimator is given by

$$\begin{aligned} x_{\mathrm{MAP}}(y) &\in \arg\max_x p_{X|Y=y}(x) \\ &= \arg\min_x \frac{1}{2\sigma^2}(y-x)^2 - \log\left(\frac{1}{2}(e^{-\frac{1}{2\varepsilon^2}(x-1)^2} + e^{-\frac{1}{2\varepsilon^2}(x+1)^2})\right) \\ &= \arg\min_x \frac{1}{2\sigma^2}(y-x)^2 + \frac{1}{2\varepsilon^2}(x^2+1) - \log\left(\cosh\left(\frac{x}{\varepsilon^2}\right)\right). \end{aligned}$$

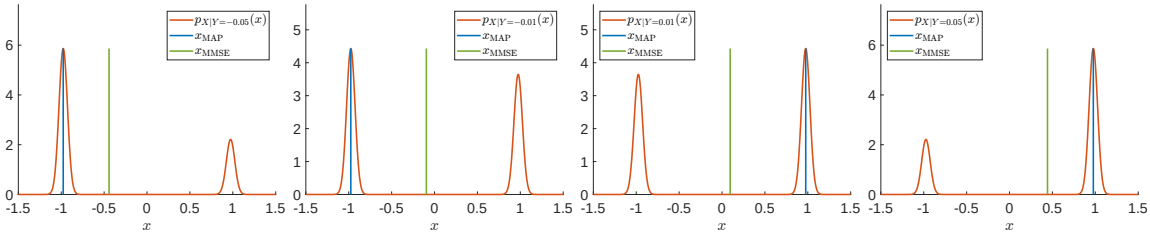
The above minimization problem has a unique global minimizer for $y \neq 0$ which we computed numerically. Figure 1 (top) shows the plot of the function $x_{\mathrm{MAP}}(y)$ for $\varepsilon^2 = 0.05^2$ and different values of σ. Clearly, small perturbations of y near zero lead to qualitatively completely different x-values, where a smaller noise level σ lowers the distance between the values $x_{\mathrm{MAP}}(y)$ for $y > 0$ and $y < 0$. In other words, the MAP estimator is not robust with respect to perturbations of the observations near zero.

In contrast, using Lemma 8, we can compute the posterior

$$P_{X|Y=y} = \frac{1}{\tilde{w}_1 + \tilde{w}_2}(\tilde{w}_1 \mathcal{N}(\cdot|\tilde{m}_1, \tilde{\sigma}^2) + \tilde{w}_2 \mathcal{N}(\cdot|\tilde{m}_2 \tilde{\sigma}^2))$$

with

$$\begin{aligned} &\tilde{\sigma}^2 = \frac{\sigma^2\varepsilon^2}{\sigma^2+\varepsilon^2}, \quad \tilde{m}_1 = \frac{\varepsilon^2 y + \sigma^2}{\varepsilon^2+\sigma^2}, \quad \tilde{m}_2 = \frac{\varepsilon^2 y - \sigma^2}{\varepsilon^2+\sigma^2}, \\ &\tilde{w}_1 = \frac{1}{2\varepsilon}\exp\left(\frac{1}{2\varepsilon^2}\left(\frac{(\varepsilon^2 y+\sigma^2)^2}{\sigma^2(\varepsilon^2+\sigma^2)} - 1\right)\right), \quad \tilde{w}_2 = \frac{1}{2\varepsilon}\exp\left(\frac{1}{2\varepsilon^2}\left(\frac{(\varepsilon^2 y-\sigma^2)^2}{\sigma^2(\varepsilon^2+\sigma^2)} - 1\right)\right). \end{aligned}$$



$y = -0.05$ $y = -0.01$ $y = 0.01$ $y = 0.05$

Figure 2: Posterior density (red), MAP estimator (blue) and MMSE estimator (green) for different observations $y = -0.05, -0.01, 0.01, 0.05$ (from left to right). While the MAP estimator is discontinuous with respect to the observation y, the posterior density is continuous with respect to y. The MMSE estimator gives just the expectation value of the posterior which is, in contrast to MAP, in not the value with highest probability.

Then the MMSE estimator is given by the expectation value of the posterior

$$\begin{aligned} x_{\text{MMSE}}(y) &= \arg\min_T \mathbb{E}_{(x,y)\sim P_{(X,Y)}} \|x - T(y)\|^2 = \mathbb{E}[X|Y=y] \\ &= \int_{\mathbb{R}} x p_{X|Y=y}(x)\,\mathrm{d}x \\ &= \frac{1}{\tilde{w}_1 + \tilde{w}_2}(\tilde{w}_1\tilde{m}_1 + \tilde{w}_2\tilde{m}_2) \\ &= \frac{1}{\tilde{w}_1 + \tilde{w}_2}\frac{1}{\varepsilon(\varepsilon^2+\sigma^2)} e^{\frac{\varepsilon^2 y^2 - \sigma^2}{2\sigma^2(\varepsilon^2+\sigma^2)}} \left(\varepsilon^2 y \cosh(\frac{y}{\varepsilon^2+\sigma^2}) + \sigma^2 \sinh(\frac{y}{\varepsilon^2+\sigma^2})\right). \end{aligned}$$

In Figure 1 (bottom), we see that the MMSE estimator shows a smooth transition in particular for larger noise levels, meaning that the estimator is robust against small perturbations of the observation near zero. The posterior is plotted for four different small values of y in Figure 2. The red curves show the graphs of the corresponding density functions. Obviously, these curves change smoothly with respect to y, i.e., we observe a continuous behavior of the posterior also with respect to observations near zero. Therefore, sampling from the posterior distributions $P_{X|Y=y}$ appears to be robust to perturbations of y. Having different samples from the posterior at $Y = y$, we can obtain a more circumvent overview on the original data than just taking their mean value represented by the MMSE estimator into account. As can be seen in the figure, for fixed y the MMSE estimator delivers just an averaged value, which is, in contrast to the non robust MAP estimator, not the one with highest probability. Note that in case of a Gaussian prior $X \sim \mathcal{N}(m, \Sigma)$ in $\mathbb{R}^m$ and white Gaussian noise, the MAP and MMSE estimators coincide.

B Local Lipschitz continuity of the generator for a latent space with infinite support

Here we show a weakened version of Lemma 1 leading to an arbitrary small additive constant. The main difference is the weaker assumption $\|\nabla_y G(y,z)\| \leq L_r$ for all $z \in \mathbb{R}^d$ with $\|z\| \leq \tilde{r}$ and all $y \in \mathbb{R}^n$ with $\|y\| \leq r$, which is fulfilled for continuously differentiable generators. For this we use the

so-called truncated normal distribution (Horrace, 2005; Tallis, 1963). Let p_Z be the density of the standard normal distribution $P_Z = \mathcal{N}(0, I_n)$, then the density of the truncated normal distribution $P_Z^{\tilde{r}}$ is given by

$$p_Z^{\tilde{r}}(z) = \begin{cases} \frac{p_Z(z)}{\int_{B_{\tilde{r}}(0)} p_Z(z)\mathrm{d}z} = \frac{p_Z(z)}{C_{\tilde{r}}}, & \text{if } \|z\| \leq \tilde{r}, \\ 0, & \text{else.} \end{cases}$$

Lemma 9. *Let $P_Z = \mathcal{N}(0, I_n)$ be the latent space. For any parameterized family of generative models G with $\|\nabla_y G(y, z)\| \leq L_r$ for all $z \in \mathbb{R}^d$ with $\|z\| \leq \tilde{r}$ and all $y \in \mathbb{R}^n$ with $\|y\| \leq r$ for some $L_r > 0$ and some $r > 0$, it holds*

$$W_1(G(y_1, \cdot)_\# P_Z, G(y_2, \cdot)_\# P_Z) \leq L_r \|y_1 - y_2\| + M_{\tilde{r}}$$

for all $y_1, y_2 \in \mathbb{R}^n$ with $\|y_1\|, \|y_2\| \leq r$. The additive constant $M_{\tilde{r}}$ fulfills $M_{\tilde{r}} \to 0$ for $\tilde{r} \to \infty$.

Proof. Let $y_1, y_2 \in \mathbb{R}^n$ with $\|y_1\|, \|y_2\| \leq r$, then it holds

$$\begin{aligned} W_1(G(y_1, \cdot)_\# P_Z, G(y_2, \cdot)_\# P_Z) \leq\ & W_1(G(y_1, \cdot)_\# P_Z, G(y_1, \cdot)_\# P_Z^{\tilde{r}}) + W_1(G(y_1, \cdot)_\# P_Z^{\tilde{r}}, G(y_2, \cdot)_\# P_Z^{\tilde{r}}) \\ & + W_1(G(y_2, \cdot)_\# P_Z^{\tilde{r}}, G(y_2, \cdot)_\# P_Z). \end{aligned}$$

By the assumption on the generator G, Lemma 1 yields

$$W_1(G(y_1, \cdot)_\# P_Z^{\tilde{r}}, G(y_2, \cdot)_\# P_Z^{\tilde{r}}) \leq L_r \|y_1 - y_2\|.$$

Consequently, it suffices to show that for $y \in \mathbb{R}^n$ with $\|y\| \leq r$ the term $W_1(G(y\,\cdot)_\# P_Z, G(y, \cdot)_\# P_Z^{\tilde{r}})$ vanishes for $\tilde{r} \to \infty$. By definition, it holds

$$\begin{aligned} W_1(G(y, \cdot)_\# P_Z, G(y, \cdot)_\# P_Z^{\tilde{r}}) &= \max_{\mathrm{Lip}(\varphi) \leq 1} \int_{\mathbb{R}^d} \varphi(G(y, z))\mathrm{d}P_Z(z) - \int_{\mathbb{R}^d} \varphi(G(y, z))\mathrm{d}P_Z^{\tilde{r}}(z) \\ &= \max_{\mathrm{Lip}(\varphi) \leq 1} \int_{\mathbb{R}^d \setminus B_{\tilde{r}}(0)} \varphi(G(y, z))\mathrm{d}P_Z(z) + \int_{B_{\tilde{r}}(0)} \varphi(G(y, z))\mathrm{d}P_Z(z) \\ &\qquad - \int_{B_{\tilde{r}}(0)} \varphi(G(y, z))\mathrm{d}P_Z^{\tilde{r}}(z) \\ &= \max_{\mathrm{Lip}(\varphi) \leq 1} \int_{\mathbb{R}^d \setminus B_{\tilde{r}}(0)} \varphi(G(y, z)) p_Z(z)\mathrm{d}z \\ &\qquad + \int_{B_{\tilde{r}}(0)} \varphi(G(y, z)) p_Z(z)(1 - \frac{1}{C_{\tilde{r}}})\mathrm{d}z. \end{aligned}$$

The first term vanishes exponentially in $\tilde{r}$ by the density p_Z, and for the second term note that $C_{\tilde{r}} \to 1$ for $\tilde{r} \to \infty$. □