SIAM Review
Vol. 65, No. 3, pp. 831–865

© 2023 Society for Industrial and Applied Mathematics

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Bayesian Inverse Problems Are Usually Well-Posed*

Jonas Latz†

Abstract. Inverse problems describe the task of blending a mathematical model with observational data—a fundamental task in many scientific and engineering disciplines. The solvability of such a task is usually classified through its well-posedness. A problem is well-posed if it has a unique solution that depends continuously on input or data. Inverse problems are usually ill-posed, but can sometimes be approached through a methodology that formulates a possibly well-posed problem. Usual methodologies are the variational and the Bayesian approach to inverse problems. For the Bayesian approach, Stuart [*Acta Numer.*, 19 (2010), pp. 451–559] has given assumptions under which the posterior measure—the Bayesian inverse problem's solution—exists, is unique, and is Lipschitz continuous with respect to the Hellinger distance and, thus, well-posed. In this work, we simplify and generalize this concept: Indeed, we show well-posedness by proving existence, uniqueness, and continuity in Hellinger distance, Wasserstein distance, and total variation distance, and with respect to weak convergence, respectively, under significantly weaker assumptions. An immense class of practically relevant Bayesian inverse problems satisfies those conditions. The conditions can often be verified without analyzing the underlying mathematical model—the model can be treated as a black box.

Key words. inverse problems, Bayesian inference, well-posedness, Kullback–Leibler divergence, total variation, Wasserstein

MSC codes. 49K40, 62F15, 65C60, 65N21, 68Q32, 68T05

DOI. 10.1137/23M1556435

Contents

*Published electronically August 8, 2023. This paper originally appeared in *SIAM/ASA Journal on Uncertainty Quantification*, Volume 8, Number 1, 2020, pages 451–482, under the title "On the Well-Posedness of Bayesian Inverse Problems."
https://doi.org/10.1137/23M1556435
Funding: The original work [54] was supported by DFG and Technische Universität München through the International Graduate School of Science and Engineering within project 10.02 BAYES.
†Maxwell Institute for Mathematical Sciences and School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK (j.latz@hw.ac.uk).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

1. Introduction. The representation of systems and processes in nature and technology with mathematical and computational models is fundamental in modern science and engineering. For a partially observable process, the *model calibration* or *inverse problem* is of particular interest. It consists of fitting model parameters such that the model represents the underlying process. Aside from classical mathematical models, such as partial differential equations and dynamical systems, inverse problems also play a central role in machine learning applications, for example, classification with deep neural networks or (non)linear support vector machines, as well as regression with deep Gaussian processes.

The solvability of inverse problems is usually classified in terms of their *well-posedness*. In 1902, Hadamard [36] defined well-posedness as follows:

. . . ces problémes . . . bien posé, je veux dire comme possible et déterminé.

In other words, according to Hadamard, a problem is *well-posed* if the solution is *possible* and *determined*; that is, it can be *found* and is *exact*. Today, we interpret this principle as follows: the solution of the inverse problem exists, is unique, and depends continuously on the data. The continuity in the data implies stability. The existence and stability allow us to find the solution (*possible*), and uniqueness makes the solution exact (*déterminé*). This is a justification for well-posedness from an

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

analytical and computational viewpoint. From a statistical viewpoint, well-posedness not only allows us to find the estimate but also gives us robustness of the estimate with respect to marginal perturbations in the data: since we often know that data is not precise, we should anticipate only marginal changes in the estimate with respect to marginal changes in the data. Otherwise, we should not consider the estimate trustworthy.

Measurement noise, complexity of the model, and a lack of data typically lead to *ill-posed* (i.e., not well-posed) inverse problems. Ill-posed inverse problems cannot be solved as they are. However, certain procedures allow us to obtain estimators for the "solution" of the ill-posed inverse problem. To be practically relevant, of course, these procedures should be well-posed. The most common methodologies are the Bayesian and the variational approaches to inverse problems. In this work, we focus on the Bayesian approach. We discuss the variational approach briefly (in subsection 2.5), but consider it generally outside of the scope of this work.

The Bayesian approach to inverse problems represents the uncertain model parameter as a random variable. The random variable is distributed according to a *prior (measure)*, which reflects uncertainty in the parameter. Observing the data is then an event with respect to which the prior shall be conditioned. The solution of the Bayesian inverse problem is this conditional probability measure, called the *posterior (measure)*. Stuart [74] transferred Hadamard's principle of well-posedness to Bayesian inverse problems: the posterior exists, it is unique, and it is locally Lipschitz continuous with respect to the data; continuity is measured in the Hellinger distance. Stuart [74] shows well-posedness under a set of sufficient but not necessary assumptions. Subsequently, Dashti and Stuart [21] reduced these assumptions significantly. Several authors have discussed what we call (*Lipschitz*, *Hellinger*) well-posedness for a variety of Bayesian inverse problems, for example, elliptic partial differential equations [20, 42], level-set inversion [43], Helmholtz source identification with Dirac sources [29], a Cahn–Hilliard model for tumor growth [47], hierarchical prior measures [56], stable priors in quasi-Banach spaces [77, 78], and convex and heavy-tailed priors [40, 41]. Finally, we mention Ernst, Sprungk, and Starkloff [30], who discussed uniform and Hölder continuity of posterior measures with respect to data and gave sufficient assumptions in this setting. We refer to these as *(Hölder, Hellinger)* and *(uniform, Hellinger) well-posedness*, respectively.

In practical applications, it may be difficult to verify (Lipschitz, Hellinger), (uniform, Hellinger), or (Hölder, Hellinger) well-posedness. The underlying mathematical model can be too complicated to analyze or may even be hidden in software. Indeed, this is the case in large-scale applications, e.g., in geotechnical engineering, meteorology, and genomics, and in machine learning algorithms. In any of these cases, the model is often a black-box, i.e., a function that takes inputs and produces deterministic outputs but with no known properties. To the best of our knowledge, it is not possible to show (Lipschitz, Hellinger) well-posedness for the Bayesian inversion of such black-box models. In turn, it may not be necessary to show (Lipschitz, Hellinger) well-posedness for many practical problems; Hadamard's concept contains only continuity, not Lipschitz continuity. In either case, we know that marginal perturbations in the data lead to marginal changes in the posterior measure. Given only continuity, the only difference is that we cannot use information about the data perturbation to quantify the change in the posterior. This, however, may be tolerable in most practical applications.

Another pressing issue is the measurement of marginal changes in the posterior. Most authors have discussed Lipschitz continuity with respect to the Hellinger distance; exceptions are, e.g., the articles of Iglesias, Lin, and Stuart [42] and Sprungk

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

[73]. While the Hellinger distance has useful properties, the actual choice of the metric should depend on the area of application.

The main contributions of this article are the following:

1. A new concept of well-posedness of Bayesian inverse problems is proposed. It consists of the existence and the uniqueness of the posterior as well as of the continuity of the data-to-posterior map in some metric or topological space of probability measures.
2. More specifically, the spaces of probability measures metrized with the Hellinger distance, the total variation distance, and the Wasserstein(p) distance, as well as those associated with the weak topology and the topology induced by the Kullback–Leibler divergence, are investigated.
3. Well-posedness of large classes of Bayesian inverse problems in any of these settings is shown. The sufficient assumptions are either nonrestrictive or easily verifiable in practical problems (e.g., when having an arbitrary model, finite-dimensional data, and nondegenerate Gaussian noise). The only actually restrictive case is that of the Kullback–Leibler topology.

This work is organized as follows. We review the Bayesian approach to inverse problems and the concept of (Lipschitz, Hellinger) well-posedness in section 2. In section 3, we advocate our relaxation of Lipschitz continuity and our consideration of metrics other than the Hellinger distance. In the same section, we introduce our notion of well-posedness and show well-posedness with respect to Hellinger, total variation, weak convergence, and the Wasserstein(p) distance, respectively. In section 4, we extend our concept to stability measurements in the Kullback–Leibler divergence, which is a quasi-semimetric. We specifically consider the case of finite-dimensional data and nondegenerate Gaussian noise in section 5. We illustrate our results numerically in section 6. In section 7, we give conclusions and point the reader towards potential future research directions. As this is a revised version of the paper [54], we also use this section for an outlook to some works that were published since publication of the original paper. Finally, we review the basics of conditional probability in Appendix A and give detailed proofs of all statements formulated in this work in Appendix B.

2. Inverse Problems, Ill-Posedness, and Well-Posedness.

2.1. Inverse Problem. We now introduce our notion of inverse problems and the necessary mathematical framework. The framework uses various concepts from measure and probability theory; for a detailed introduction, we recommend, e.g., the book by Ash and Doléans-Dade [2] or the book by Billingsley [7].

Let $y^\dagger$ be *(observational) data* in some separable Banach space $(Y, \|\cdot\|_Y)$; this is the *data space*. The data shall be used to train a mathematical model, that is, identify a *(model) parameter* $\theta^\dagger$ in a set X. The *parameter space* X is a measurable subset of some Radon space $(X', \mathcal{T}')$; i.e., $(X', \mathcal{T}')$ is a separable, completely metrizable topological vector space. X' could, for instance, also be a separable Banach space. Moreover, X and Y form measurable spaces with their respective Borel-σ-algebras $\mathcal{B}X := \mathcal{B}(X, X \cap \mathcal{T}')$ and $\mathcal{B}Y := \mathcal{B}(Y, \|\cdot\|_Y)$. Let $\mathcal{G} : X \to Y$ be a measurable function called the *forward response operator*. It represents the connection between parameter and data in the mathematical model. We define the inverse problem as follows:

$$\text{(IP)} \qquad \text{Find } \theta^\dagger \in X, \text{ such that } y^\dagger = \mathcal{G}(\theta^\dagger) + \eta^\dagger.$$

Here, $\eta^\dagger \in Y$ is *(observational) noise*. The given "additive noise" setting is usual in

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

many practical applications. We mainly use it here to discuss some basic principles; many of the results shown throughout the article do not actually rely on this structure.

We discuss the solvability and stability of inverse problems in terms of well-posedness.

DEFINITION 2.1 (well-posedness). *The problem* (IP) *is* well-posed *if*

1. *this problem has a solution* (existence),
2. *the solution is unique* (uniqueness)*, and*
3. *the solution depends continuously on the data* y (stability).

A problem that is not well-posed is called ill-posed.

We generally consider the observational noise $\eta^\dagger$ to be unknown and model it as a realization of a random variable $\eta \sim \mu_{\text{noise}}$. If the noise takes any value in Y, the problem (IP) is ill-posed.

PROPOSITION 2.2. *Let* X *contain at least two elements, and let the support of* μ_{noise} *be* Y*. Then the inverse problem* (IP) *is ill-posed.*

Note that the assumptions in Proposition 2.2 can often be verified. If X contains only one element, the inverse problem is uniquely solvable. However, there is only one possible parameter $\theta \in X$, which makes the inverse problem trivial. If Y is finite-dimensional, the second assumption is, for instance, fulfilled when μ_{noise} is nondegenerate Gaussian.

2.2. Bayesian Inverse Problem. The Bayesian approach to (IP) proceeds as follows. First, we model the parameter $\theta \sim \mu_{\text{prior}}$ as a random variable. μ_{prior} is the so-called prior measure. This probability measure reflects the uncertainty in the parameter.[1] Moreover, we assume that θ, η are independent random variables defined on an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$. In this setting, the inverse problem (IP) is an event,

$$\{y^\dagger = \mathcal{G}(\theta) + \eta\} \in \mathcal{A},$$

where the data $y^\dagger$ is a realization of the random variable $\mathcal{G}(\theta^\dagger) + \eta$. The solution to the Bayesian inverse problem is the posterior measure

$$\mu^\dagger_{\text{post}} := \mathbb{P}(\theta \in \cdot | \mathcal{G}(\theta) + \eta = y^\dagger). \tag{2.1}$$

For the definition, existence, and uniqueness statement concerning conditional probabilities, we refer the reader to Appendix A.

First, note that we define $y := \mathcal{G}(\theta) + \eta$ to be a random variable reflecting the distribution of the data, given an uncertain parameter. We can deduce the conditional measure of the data y given $\theta = \theta'$:

$$\mu_L = \mathbb{P}(y \in \cdot | \theta = \theta') = \mu_{\text{noise}}(\cdot - \mathcal{G}(\theta')).$$

To this end, note again that the inverse problem setting $y^\dagger = \mathcal{G}(\theta) + \eta$ is only a specific example. In the following, we consider more general Bayesian inverse problems. Now, y is a random variable on $(Y, \mathcal{B}Y)$ depending on θ. The conditional probability of y given that $\theta = \theta'$ is defined by μ_L, which now fully describes the dependence of θ and y. In case it exists, the forward response operator $\mathcal{G}$ is implicitly part of μ_L.

[1] The use of probabilities to reflect uncertainties, rather than (only) randomness, has been discussed by, e.g., Bayes [4], Cox [17], and Jaynes [45] and opposes the classical view of, e.g., Kolmogorov [51] and von Mises [81]. Schwarz [71] criticizes both (or all) interpretations of probability.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Remark 2.3. From a statistical viewpoint, we consider a parametric statistical model $(Y, \mathcal{P})$, where $\mathcal{P} := \{\mu_L(\cdot|\theta') : \theta' \in X\}$. Hence, the data $y^\dagger \in Y$ is a realization of $y \sim \mu_L(\cdot|\theta^\dagger)$ for $\theta^\dagger \in X$. The data $y^\dagger$ is then used to identify this $\theta^\dagger$ from among the other elements of X. For a thorough discussion of statistical models, we refer the reader to [64]. We note that throughout this work, we use the denomination "parametric," even if X is an infinite-dimensional space—as opposed to the usual nomenclature in the statistical literature, which uses the term "parametric" only in the case where X is finite-dimensional, e.g., [16, 33].

Given μ_{prior} and μ_L, we apply Bayes' theorem to find the posterior measure $\mu^\dagger_{\text{post}}$, now given by

$$\mu^\dagger_{\text{post}} := \mathbb{P}(\theta \in \cdot | y = y^\dagger). \tag{2.2}$$

Bayes' theorem gives a connection of $\mu_{\text{prior}}, \mu^\dagger_{\text{post}}$, and μ_L in terms of their *probability density functions* (pdfs). We obtain these pdfs by assuming that there are σ-finite measure spaces $(X, \mathcal{B}X, \nu_X)$ and $(Y, \mathcal{B}Y, \nu_Y)$, where $\mu_{\text{prior}} \ll \nu_X$ and $\mu_L(\cdot|\theta') \ll \nu_Y$ for $\theta' \in X$, μ_{prior}-almost surely (a.s.). The Radon–Nikodym theorem implies that the following pdfs exist:

$$\frac{\mathrm{d}\mu_L}{\mathrm{d}\nu_Y}(y^\dagger) =: L(y^\dagger|\theta'), \qquad \frac{\mathrm{d}\mu_{\text{prior}}}{\mathrm{d}\nu_X}(\theta) =: \pi_{\text{prior}}(\theta).$$

The conditional density $L(\cdot|\theta')$ is called *(data) likelihood*. The dominating measures ν_X, ν_Y are often (but not exclusively) given by the counting measure, the Lebesgue measure, or a Gaussian measure. For example, if X is infinite-dimensional and μ_{prior} is Gaussian, we set $\nu_X := \mu_{\text{prior}}$ and $\pi_{\text{prior}} \equiv 1$. The posterior measure is then given in terms of a pdf with respect to the Gaussian prior measure. This setting is thoroughly discussed in [21, 74]; however, it is also contained in our version of Bayes' theorem. Before moving on to that, we discuss a measure-theoretic subtlety we encounter with conditional probabilities and their densities.

Remark 2.4. Conditional probabilities such as $\mu^\dagger_{\text{post}} = \mathbb{P}(\theta \in \cdot | y = y^\dagger)$ are uniquely defined only for $\mathbb{P}(y \in \cdot)$-almost every (a.e.) $y^\dagger \in Y$; see Theorem A.1. This implies that if $\mathbb{P}(y \in \cdot)$ has a continuous distribution, point evaluations in Y of the function $\mathbb{P}(\theta \in A | y = \cdot)$ may not be well defined for $A \in \mathcal{B}X$. In this case, one would not be able to compute the posterior measure for any single-point data set $y^\dagger \in Y$. Also, the statements (2.1), (2.2), as well as the definition of the likelihood, should be understood only for $\mathbb{P}(y \in \cdot)$-a.e. $y^\dagger \in Y$.

Our version of Bayes' theorem is mainly built on [21, Theorem 3.4]. However, in the proof we need neither assume that the model evidence is positive and finite nor assume continuity in the data or the parameter of the likelihood.

THEOREM 2.5 (Bayes). *Let $y^\dagger \in Y$ be $\mathbb{P}(y \in \cdot)$-a.s. defined. Moreover, let $L(y^\dagger|\cdot)$ be integrable, i.e., in $\mathbf{L}^1(X, \mu_{\text{prior}})$, and strictly positive. Then,*

$$Z(y^\dagger) := \int L(y^\dagger|\theta) \mathrm{d}\mu_{\text{prior}}(\theta) \in (0, \infty).$$

Moreover, the posterior measure $\mu^\dagger_{\text{post}} \ll \nu_X$ exists, it is unique, and it has the ν_X-density

$$\pi^\dagger_{\text{post}}(\theta') = \frac{L(y^\dagger|\theta')\pi_{\text{prior}}(\theta')}{Z(y^\dagger)} \qquad (\theta' \in X, \nu_X\text{-a.s.}). \tag{2.3}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

The quantity in the denominator of Bayes' formula $Z(y^\dagger) := \int L(y^\dagger|\theta)\mathrm{d}\mu_{\mathrm{prior}}(\theta)$ is the ν_Y-density of $\mathbb{P}(y \in \cdot)$ and is called *(model) evidence*. We comment on the assumptions of Theorem 2.5 in subsection 3.4. In Remark 2.4, we mention that the posterior measure is only $\mathbb{P}(y \in \cdot)$-a.s. uniquely defined. Hence, the map $y^\dagger \mapsto \mu^\dagger_{\mathrm{post}}$ is not well defined. We resolve this issue by fixing the definition of the likelihood $L(y^\dagger|\theta')$ for every $y^\dagger \in Y$ and μ_{prior}-a.e. $\theta' \in X$. According to Theorem 2.5, we then indeed obtain a unique posterior measure for any data set $y^\dagger \in Y$. We define the *Bayesian inverse problem* with prior μ_{prior} and likelihood L by

$$\text{(BIP)} \qquad \text{Find } \mu^\dagger_{\mathrm{post}} \in \mathrm{Prob}(X, \mu_{\mathrm{prior}}) \text{ with } \nu_X\text{-density } \pi^\dagger_{\mathrm{post}}(\theta) = \frac{L(y^\dagger|\theta)\pi_{\mathrm{prior}}(\theta)}{Z(y^\dagger)}.$$

Here, $\mathrm{Prob}(X, \mu_{\mathrm{prior}})$ denotes the set of probability measures on $(X, \mathcal{B}X)$ which are absolutely continuous with respect to the prior μ_{prior}. Similarly, we define the set of all probability measures on $(X, \mathcal{B}X)$ by $\mathrm{Prob}(X)$. If X forms a normed space with some norm $\|\cdot\|_X$, we define the set of probability measures with finite pth moment by

$$\mathrm{Prob}_p(X) := \left\{ \mu \in \mathrm{Prob}(X) : \int \|\theta\|_X^p \mu(\mathrm{d}\theta) < \infty \right\} \qquad (p \in [1, \infty)).$$

2.3. Degenerate Bayesian Inverse Problems. There are Bayesian inverse problems for which Bayes' theorem (Theorem 2.5) is not satisfied. Consider $\mu_{\mathrm{noise}} := \delta(\cdot - 0)$ as a noise distribution; i.e., the noise is a.s. 0. We refer to Bayesian inverse problems with such a noise distribution as *degenerate*, since the noise distribution is degenerate. Here, we represent the likelihood by

$$L(y^\dagger|\theta') := \begin{cases} 1 & \text{if } y^\dagger = \mathcal{G}(\theta'), \\ 0 & \text{otherwise.} \end{cases}$$

Due to different dimensionality, it is now likely that the prior μ_{prior} is chosen such that it gives probability 0 to the solution manifold $S = \{\theta' \in X : y^\dagger = \mathcal{G}(\theta')\}$, i.e., $\mu_{\mathrm{prior}}(S) = 0$. Then, we have

$$Z(y^\dagger) = \int_X L(y^\dagger|\theta)\mu_{\mathrm{prior}}(\mathrm{d}\theta) = \int_S 1\mu_{\mathrm{prior}}(\mathrm{d}\theta) = \mu_{\mathrm{prior}}(S) = 0$$

and do not obtain a valid posterior measure for $y^\dagger$ from Theorem 2.5. Alternatively, we can employ the disintegration theorem; see Cockayne et al. [14] and the definition of conditional probabilities after Theorem A.1. In the following proposition, we give a simple example for such (BIP).

PROPOSITION 2.6. *Let $\mathcal{G} : X \to Y$ be a homeomorphism; i.e., it is continuous and bijective, and $\mathcal{G}^{-1} : Y \to X$ is continuous as well. Moreover, let $\mu_{\mathrm{prior}} \in \mathrm{Prob}(X)$ be some prior measure, and let $\mu_{\mathrm{noise}} = \delta(\cdot - 0)$. Then, $\mu^\dagger_{\mathrm{post}} = \delta(\mathcal{G}(\cdot) - y^\dagger) = \delta(\cdot - \mathcal{G}^{-1}(y^\dagger))$ for $\mu_{\mathrm{prior}}(\mathcal{G} \in \cdot)$-a.e. $y^\dagger \in Y$.*

Note that we cannot easily solve the problem discussed in Remark 2.4 for this Bayesian inverse problem. Hence, point evaluations $y^\dagger \mapsto \mu^\dagger_{\mathrm{post}}$ may indeed not be well defined in this setting. Therefore, when discussing this Bayesian inverse problem, we will fix one representative in the class of measures that are a.s. equal to the posterior.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

2.4. Lipschitz Well-Posedness. We now move on to the definition of Stuart's [74] concept of well-posedness of Bayesian inverse problems. Similarly to the well-posedness definition of the classical problem (IP), we consider an existence, a uniqueness, and a stability condition; see Definition 2.1. Stability is quantified in terms of the *Hellinger distance*

$$\mathrm{d}_{\mathrm{Hel}}(\mu,\mu') = \sqrt{\frac{1}{2}\int \left(\sqrt{\frac{\mathrm{d}\mu'}{\mathrm{d}\mu_{\mathrm{prior}}}} - \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\mu_{\mathrm{prior}}}}\right)^2 \mathrm{d}\mu_{\mathrm{prior}}}$$

between two measures $\mu, \mu' \in \mathrm{Prob}(X, \mu_{\mathrm{prior}})$. The Hellinger distance is based on the work [38]. With this, we can now formalize the concept of (Lipschitz, Hellinger) well-posedness for Bayesian inverse problems.

DEFINITION 2.7 ((Lipschitz, Hellinger) well-posedness). *The problem* (BIP) *is* (Lipschitz, Hellinger) well-posed *if*

1. $\mu^\dagger_{\mathrm{post}} \in \mathrm{Prob}(X, \mu_{\mathrm{prior}})$ *exists* (existence),
2. $\mu^\dagger_{\mathrm{post}}$ *is unique in* $\mathrm{Prob}(X, \mu_{\mathrm{prior}})$ (uniqueness)*, and*
3. $(Y, \|\cdot\|_Y) \ni y^\dagger \mapsto \mu^\dagger_{\mathrm{post}} \in (\mathrm{Prob}(X, \mu_{\mathrm{prior}}), \mathrm{d}_{\mathrm{Hel}})$ *is locally Lipschitz continuous* (stability).

2.5. Variational Inverse Problems. We have presented the Bayesian approach to inverse problems here as a natural way to formulate and solve the inverse problem (IP) from a probabilistic perspective. Although it does not appear in the remainder of the article, for the sake of completeness, we briefly comment now on the *variational approach to inverse problems*. For more details, we refer to, e.g., [11, 12, 34]. The fundamental idea traces back to classical *maximum likelihood estimation*: we aim to find the parameter that maximizes the likelihood function given the observed data $y^\dagger$. In particular, we find

$$\theta_{\mathrm{ml}} \in \mathrm{argmin}_{\theta' \in X} - \log L(y^\dagger|\theta').$$

This problem is, for instance, a (possibly nonlinear) least-squares problem if μ_{noise} is nondegenerate Gaussian. This problem is still likely to be ill-posed: the optimization problem may be nonconvex or the parameter underdetermined, which in both cases can lead to multiple global minimizers. Even if there is a single global minimizer, discontinuity of $y^\dagger \mapsto \theta_{\mathrm{ml}}$ appears in, e.g., image deblurring, where $\mathcal{G}$ is often linear and invertible, but not boundedly invertible [34]. Another example for the ill-posedness of maximum likelihood estimation is given by [48] in the context of estimating hyperparameters in Gaussian process regression.

The ill-posedness of the variational inverse problem can often be mitigated through appropriate *regularization*, e.g., solving

$$\theta_{\mathrm{reg}} \in \mathrm{argmin}_{\theta' \in X} - \log L(y^\dagger|\theta') + R(\theta')$$

for an appropriate function $R : X \to \mathbb{R}$, the *regularizer*. R is used to introduce additional information about $\theta^\dagger$, e.g., that it lives in a certain subspace of X, that it is close to some value $\theta_0 \in X$, or that it is sparse, i.e., it is a vector that has many zero entries. Some examples for well-posed regularized variational inverse problems can be found in [13].

The regularizer is conceptually very similar to a prior measure. Indeed, in finite-dimensional settings, where $\pi^\dagger_{\mathrm{post}}$ is a Lebesgue density, the *maximum a posteriori* (MAP) estimator

$$\theta_{\mathrm{MAP}} \in \mathrm{argmin}_{\theta' \in X} - \log \pi^\dagger_{\mathrm{post}}(\theta')$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

corresponds to a regularized variational inverse problem, with $R = -\log \pi_{\text{prior}}$. In the infinite-dimensional setting a similar correspondence of MAP and variational inverse problem can often be shown to hold true. As there is no Lebesgue density in infinite dimensions, we need a different way to represent the MAP estimator. A natural way is to consider the $\theta' \in X$ that maximizes the limit

$$\lim_{\delta \downarrow 0} \frac{\mu^{\dagger}_{\text{post}}(B(\theta', \delta))}{\mu^{\dagger}_{\text{post}}(B(\theta'', \delta))}$$

for all $\theta'' \in X$ and where $B(\theta'', \delta) \in \mathcal{B}X$ denotes the open ball with center $\theta'' \in X$ and radius $\delta > 0$. MAP estimators in this infinite-dimensional setting have been studied in, e.g., [19, 37, 60], where they can indeed be determined through a regularized variational inverse problem. The opposite correspondence between MAP and regularized variational inverse problems is not always true. Indeed, $\exp(-R)$ does not need to represent a prior measure, e.g., [53].

3. Redefining Bayesian Well-Posedness. In this work, we try to identify general settings in which we can show some kind of well-posedness of (BIP), using no or very limited assumptions on the underlying mathematical model or the forward response operator. In particular, we aim to find verifiable assumptions on the likelihood $L(y^{\dagger}|\theta')$ (or rather the noise model) that are independent of the underlying forward response operator

$$\mathcal{G} \in \mathbf{M} := \{f : X \to Y \text{ measurable}\}.$$

Neglecting Proposition 2.6 for a moment, existence and uniqueness are often implied by Theorem 2.5. However, the local Lipschitz continuity condition, reflecting stability, is rather strong. In subsection 3.1, we give examples in which local Lipschitz continuity does not hold in the posterior measure or is hard to verify by using results in the literature. In any of these cases, we show that the posterior measures are continuous in the data. Given that the classical formulation of well-posedness, i.e., Definition 2.1, does not require local Lipschitz continuity and that local Lipschitz continuity may be too strong for general statements, we use these examples to advocate a relaxation of the local Lipschitz continuity condition.

Moreover, it is not possible to use the Hellinger distance to quantify the distance between two posteriors in some situations. In other situations, the Hellinger distance may be inappropriate from a contextual viewpoint. In subsection 3.2, we will investigate these issues as motivation to consider metrics other than the Hellinger distance.

In subsection 3.3, we will introduce the concept of (P, d)-well-posedness of Bayesian inverse problems. Finally, we will show the main results of this work: we give conditions under which we can show well-posedness in a variety of metrics in subsections 3.4 and 3.5.

3.1. Relaxing the Lipschitz Condition. Ill-posedness in the (Lipschitz, Hellinger) sense can, for instance, occur when data has been transformed by a non-Lipschitz continuous function. As an example, we consider a Bayesian inverse problem that is linear and Gaussian; however, the data is transformed by the cube root function.

Example 3.1. Let $X := Y := \mathbb{R}$. We consider the Bayesian approach to the inverse problem

$$y^{\dagger} = (\theta + \eta)^3,$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

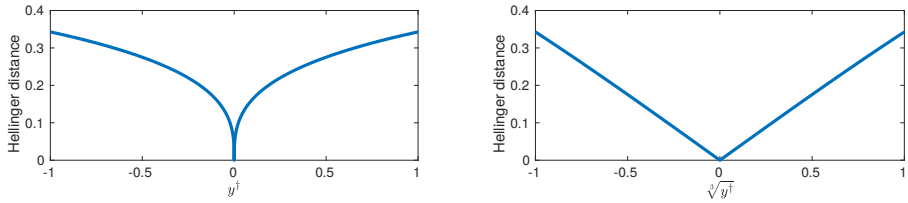

Fig. 1 *Hellinger distances between posterior measures in Example* 3.1. *The posterior measures are based on two data sets: $y^\dagger$ that varies in $(-1,1)$ and $y^\ddagger := 0$. In the left figure, we show the relationship between the data and the Hellinger distance. In the right figure, we replace the data by $y^\dagger := \sqrt[3]{y^\dagger}$, $y^\ddagger := \sqrt[3]{y^\ddagger}$. In both plots, we observe a continuous relationship between the Hellinger distance and the data, which is also Lipschitz continuous in the right figure but not in the left figure.*

where θ is the unknown parameter and η is observational noise; both are independent. The probability measures of the parameter and the noise are given by $\mu_{\text{prior}} := \mu_{\text{noise}} := \mathrm{N}(0, 1^2)$. The likelihood of (BIP) is

$$L(\theta|y^\dagger) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\|\theta - \sqrt[3]{y^\dagger}\|^2\right).$$

Since both the prior and noise are Gaussian, and the forward model is linear (the identity operator), we can compute the posterior measure analytically; see [1, section 3]. We obtain $\mu^\dagger_{\text{post}} := \mathrm{N}(\sqrt[3]{y^\dagger}/2, (1/\sqrt{2})^2)$. Moreover, one can show that

$$\mathrm{d}_{\text{Hel}}(\mu^\dagger_{\text{post}}, \mu^\ddagger_{\text{post}}) = \sqrt{1 - \exp\left(-\frac{1}{8}\left(\sqrt[3]{y^\dagger} - \sqrt[3]{y^\ddagger}\right)^2\right)}, \tag{3.1}$$

where $\mu^\ddagger_{\text{post}}$ is the posterior measure based on a second data set $y^\ddagger \in Y$. One can show analytically that this Hellinger distance in (3.1) is not locally Lipschitz as $|y^\dagger - y^\ddagger| \to 0$. It is, however, continuous. We plot the Hellinger distance in Figure 1 on the left-hand side, where we set $y^\ddagger := 0$ and vary only $y^\dagger \in (-1, 1)$. We observe indeed that the Hellinger distance is continuous but not Lipschitz continuous. In the plot on the right-hand side, we show the Hellinger distance when considering $\sqrt[3]{y^\dagger}$, rather than $y^\dagger$, as the data set. In this case, the Hellinger distance is locally Lipschitz in the data.

The Bayesian inverse problem in Example 3.1 is ill-posed in the sense of Definition 2.7, since the posterior is only continuous but not Lipschitz in the data. However, we can heal this ill-posedness by transforming $y^\dagger \mapsto \sqrt[3]{y^\dagger}$. Hence, the (Lipschitz, Hellinger) well-posedness property reduces to a continuous data transformation problem.

Other examples may be (Lipschitz, Hellinger) well-posed, but this may be difficult to verify in practice or for general forward response operators. Dashti and Stuart give [21, Assumptions 1] that are sufficient, but not necessary, to prove well-posedness. One of the assumptions is local Lipschitz continuity in the log-likelihood $\log L$ with respect to the data. Here, the Lipschitz constant is assumed to be a positive function that is monotonically nondecreasing in $\|\theta\|_X$. This assumption is not satisfied in the following example.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Example 3.2. Let $X := (0,1)$ and $Y := \mathbb{R}$. We consider the Bayesian approach to the inverse problem

$$y^\dagger = \theta^{-1} + \eta,$$

where θ is the unknown parameter and η is observational noise. Neglecting linear prefactors, this inverse problem can be thought of as the recovery of a wavelength θ from a noisy frequency measurement $y^\dagger$.

The prior measure of θ is given by $\mu_{\text{prior}} = \text{Unif}(0,1)$. The noise is distributed according to $\mu_{\text{noise}} = \text{N}(0, 1^2)$. Moreover, note that both the parameter and noise are independent random variables. The likelihood of (BIP) is

$$L(y^\dagger|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\|\theta^{-1} - y^\dagger\|^2\right).$$

For fixed $\theta \in X$, the logarithm of the likelihood in this setting is Lipschitz continuous in the data. However, as $\theta \downarrow 0$, the Lipschitz constant explodes. Hence, the likelihood does not fulfill [21, Assumptions 1].

Hence, we cannot use the theory of Dashti and Stuart [21, section 4] to show (Lipschitz, Hellinger) well-posedness of the Bayesian inverse problem in Example 3.2. We expect a similar problem for forward response operators that are not locally bounded. In Corollary 5.1, we revisit Example 3.2 and show that the posterior measure is continuous with respect to the data.

Up to now we have presented academic examples. A practical and more relevant problem is the Bayesian elliptic inverse problem. It is the prototype example in the context of partial differential equations and has been investigated by various authors, e.g., [20, 21, 56, 67, 74].

Example 3.3 (elliptic inverse problem). Let the parameter space be the space of continuously differentiable functions $X := \mathbf{C}^1(D)$ on a connected, bounded open set $D \subseteq \mathbb{R}^d$, $d = 1, 2, 3$, with smooth boundary. The data space $Y := \mathbb{R}^k$ is finite-dimensional. Moreover, let $f \in \mathbf{C}^0(D)$ be a continuous function. The underlying model is an elliptic partial differential equation,

$$\begin{aligned} -\nabla \cdot \left(\mathrm{e}^{\theta(x)} \nabla p(x)\right) &= f(x) && (x \in D), \\ p(x) &= 0 && (x \in \partial D), \end{aligned}$$

or rather its weak formulation. In a typical application, the solution p represents the pressure head in a groundwater reservoir, while the diffusion coefficient $\exp(\theta(x))$ represents the reservoir's hydraulic conductivity. Noisy measurements of the pressure head at locations $x_1, \ldots, x_k \in D$ shall now be used to infer the log-conductivity θ. Hence, the forward response operator is the map

$$\mathcal{G} : X \to Y, \qquad \theta \mapsto (p(x_1), \ldots, p(x_k)).$$

In practical applications, allowing only continuously differentiable functions to be diffusion coefficients may be too restrictive. Iglesias, Lin, and Stuart [42] consider more realistic geometric prior measures. In [42, Theorem 3.5], the authors show local Lipschitz continuity for some of those prior measures, but they show only Hölder continuity with coefficient $\gamma = 0.5$ for others. This is another example where (Lipschitz, Hellinger) well-posedness in the sense of Definition 2.7 has not been shown, but continuity in the posterior measure is satisfied.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

In the following, in subsection 3.3, we weaken the *Lipschitz well-posedness* by replacing Lipschitz continuity with continuity as a stability condition. Looking back at Examples 3.1–3.3, we consider this weakening to be tolerable for practical problems.

3.2. Reconsidering the Hellinger Distance. The Hellinger distance is a popular choice for analyzing the continuous dependence or, e.g., the approximation of measures. However, there are cases in which it cannot be used.

We consider the Bayesian inverse problem discussed in Proposition 2.6. We set $\mu^\dagger_{\text{post}} := \delta(\cdot - \mathcal{G}^{-1}(y^\dagger))$ as a posterior measure with $\mathcal{G}^{-1} : Y \to X$ continuous. We set $X := Y := \mathbb{R}$ and $\mu_{\text{prior}} := \mathrm{N}(0, 1^2)$. Then $\mu^\dagger_{\text{post}} \not\ll \mu_{\text{prior}}$ for $y^\dagger \in Y$. The Hellinger distance between $\mu^\dagger_{\text{post}}$ and $\mu^\ddagger_{\text{post}}$ is not well defined for any other data set $y^\ddagger \neq y^\dagger$. Instead, we consider the closely related total variation (tv) distance and obtain

$$\mathrm{d_{tv}}(\mu^\dagger_{\text{post}}, \mu^\ddagger_{\text{post}}) := \sup_{B \in \mathcal{B}X} \left| \mu^\dagger_{\text{post}}(B) - \mu^\ddagger_{\text{post}}(B) \right| = 1.$$

Hence, $\mu^\ddagger_{\text{post}} \not\to \mu^\dagger_{\text{post}}$ in tv as $y^\ddagger \to y^\dagger$. Thus, the Bayesian inverse problem is not stable in the tv distance, i.e., ill-posed in this sense.

However, we have $\mu^\ddagger_{\text{post}} \to \mu^\dagger_{\text{post}}$ weakly as $y^\ddagger \to y^\dagger$. Hence, we observe continuity in the weak topology on the space $\mathrm{Prob}(X)$ of probability measures on $(X, \mathcal{B}X)$. Equivalently, we can say that we observe continuity in the *Prokhorov metric* on $\mathrm{Prob}(X)$,

$$\mathrm{d_{Prok}}(\mu, \mu') := \inf \left\{ \varepsilon > 0 : \mu(B) \leq \mu'(B^\varepsilon) + \varepsilon, B \in \mathcal{B}X \right\},$$

where $B^\varepsilon := \{b \in X : b' \in B, \|b - b'\|_X < \varepsilon\}$ is the open generalized ε-ball around $B \in \mathcal{B}X$; see [66] for details.

To summarize, there are cases in which the Hellinger distance is infeasible for showing well-posedness. Moreover, different metrics may lead to different well-posedness results. Hence, we should introduce a concept that allows for different metrics on the space of probability measures.

3.3. Definition. As motivated in subsections 3.1 and 3.2, we next generalize the notion of well-posedness of Bayesian inverse problems. In Definition 2.7, we considered Lipschitz continuity in the Hellinger distance as a stability criterion. Now, we consider simple continuity with respect to various metric spaces.

DEFINITION 3.4 ((P, d)-well-posedness). *Let* (P, d) *be a metric space of probability measures on* $(X, \mathcal{B}X)$*, i.e.,* $\mathrm{P} \subseteq \mathrm{Prob}(X)$*. The problem* (BIP) *is* (P, d)-well-posed *if*

1. $\mu^\dagger_{\text{post}} \in \mathrm{P}$ *exists* (existence);
2. $\mu^\dagger_{\text{post}}$ *is unique in* P (uniqueness); *and*
3. $(Y, \|\cdot\|_Y) \ni y^\dagger \mapsto \mu^\dagger_{\text{post}} \in (\mathrm{P}, d)$ *is a continuous function* (stability).

For particular (P, d), we introduce special denominations. Indeed, we denote (P, d)-well-posedness by

(i) *weak* well-posedness if we consider the *Prokhorov metric*; i.e., we set $(\mathrm{P}, d) = (\mathrm{Prob}(X), \mathrm{d_{Prok}})$;
(ii) *tv* well-posedness if we consider the tv; i.e., we set $(\mathrm{P}, d) := (\mathrm{Prob}(X), \mathrm{d_{tv}})$;
(iii) *Hellinger* well-posedness if we consider the *Hellinger distance;* i.e., we set $(\mathrm{P}, d) := (\mathrm{Prob}(X, \mu_{\text{prior}}), \mathrm{d_{Hel}})$; and
(iv) *Wasserstein(p)* well-posedness if X is a normed space and if we consider the *Wasserstein(p) distance*; i.e., we set $(\mathrm{P}, d) := (\mathrm{Prob}_p(X), \mathrm{d_{Was(p)}})$ for some $p \in [1, \infty)$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Which concept of well-posedness should we consider in practice? Weak well-posedness implies continuity of posterior expectations of bounded, continuous quantities of interest. If this is the task of interest, weak well-posedness should be sufficient. Hellinger and tv distance imply convergence of the posterior expectation of any bounded quantity of interest. Hence, if discontinuous functions shall be integrated, or probabilities computed, those distances should be chosen. Wasserstein(p) distances have gained popularity in the convergence and stability theory of Markov chain Monte Carlo (MCMC) algorithms; see, e.g., [31, 69]. Hence, Wasserstein(p) well-posedness may be the right tool when discussing the well-posedness of solving a Bayesian inverse problem via MCMC.

3.4. Hellinger, Total Variation, and Weak Well-Posedness. We now give assumptions under which a Bayesian inverse problem can be shown to be Hellinger well-posed, tv well-posed, and weakly well-posed.

ASSUMPTIONS 3.5. *Consider* (BIP). *Let the following assumptions hold for* μ_{prior}*-a.e.* $\theta' \in X$ *and every* $y^\dagger \in Y$.

(A1) $L(\cdot|\theta')$ *is a strictly positive probability density function;*
(A2) $L(y^\dagger|\cdot) \in \mathbf{L}^1(X, \mu_{\text{prior}})$*;*
(A3) $g \in \mathbf{L}^1(X, \mu_{\text{prior}})$ *exists such that* $L(y^\ddagger|\cdot) \leq g$ *for all* $y^\ddagger \in Y$*; and*
(A4) $L(\cdot|\theta')$ *is continuous.*

(A1) means that any data set $y^\dagger \in Y$ has a positive likelihood under any parameter $\theta' \in X$. We conservatively assume that no combination of parameter and data values is impossible, but some may be unlikely. This can usually be satisfied by continuously transforming the forward response operator and/or by choosing a noise distribution that is concentrated on all of Y. Note that the assumption that $L(y^\dagger|\theta')$ is a pdf can be relaxed to $c \cdot L(y^\dagger|\theta')$ being a pdf, where $c > 0$ depends on neither $y^\dagger$ nor θ'. (A2)–(A3) imply that the likelihood is integrable with respect to the prior and that it is bounded from above uniformly in the data by an integrable function. These assumptions are satisfied, for instance, when the likelihood is bounded from above by a constant. Noise models with bounded pdfs on Y should generally imply a bounded likelihood. (A4) requires the continuity of the likelihood with respect to the data. Continuity in the data is given, for instance, when considering noise models with continuous pdfs and a continuous connection between the noise and the model. We give examples in section 6 showing that we cannot neglect the continuity in the data. Here, we show Hellinger, tv, and weak well-posedness under (A1)–(A4).

THEOREM 3.6. *Let* (A1)–(A4) *hold for* (BIP). *Then* (BIP) *is weakly, Hellinger, and total variation (whtv) well-posed.*

For the proof of this theorem, we proceed as follows. First, we show that (A1)–(A4) imply Hellinger well-posedness. Then we show that tv well-posedness and weak well-posedness are indeed implied by Hellinger well-posedness by some topological argument.

LEMMA 3.7. *Let* (A1)–(A4) *hold for* (BIP). *Then* (BIP) *is Hellinger well-posed.*

We can bound Prokhorov and tv distance with the Hellinger distance; see [32] for the appropriate results. In such a case, the continuity of a function in the bounding metric immediately implies continuity also in the bounded metric.

LEMMA 3.8. *Let* A, B *be two sets, and let* (A, d_A), (B, d_1), *and* (B, d_2) *be metric spaces. Let* $f : (A, d_A) \to (B, d_2)$ *be a continuous function. Moreover, let* $t : [0, \infty) \to$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$[0,\infty)$ be continuous in 0, with $t(0)=0$. Finally, let

$$d_1(b,b') \leq t(d_2(b,b')) \qquad (b,b' \in B).$$

Then $f:(A,d_A)\to(B,d_1)$ is continuous as well.

In the setting of Lemma 3.8, we call d_1 *coarser* than d_2, respectively, d_2 *finer* than d_1. The lemma implies that if we are going from a finer to a coarser metric, continuous functions keep being continuous. By the bounds given in [32], Prokhorov and tv distance are coarser than the Hellinger distance. If the function $y^\dagger \mapsto \mu^\dagger_{\text{post}}$ is continuous in the Hellinger distance, it is also continuous in the weak topology and the tv distance. We summarize this result in the following proposition.

PROPOSITION 3.9. *Let d_1, d_2 be metrics on* P, *and let d_1 be coarser than d_2. Then, a Bayesian inverse problem that is* (P, d_2)*-well-posed is also* (P, d_1)*-well-posed.*

Therefore, Hellinger well-posedness (in Lemma 3.7) implies tv and weak well-posedness (in Theorem 3.6).

3.5. Wasserstein(p) Well-Posedness. Let $p\in[1,\infty)$, and let X form a normed space with norm $\|\cdot\|_X$. The *Wasserstein(p) distance* between $\mu,\mu'\in \mathrm{Prob}_p(X)$ can be motivated by the theory of optimal transport. It is given as the cost of the optimal transport from μ to μ'. The cost of transport from $\theta\in X$ to $\theta'\in X$ is given by $\|\theta-\theta'\|_X$. More precisely, the *Wasserstein(p)* distance (i.e., the Wasserstein distance of order p) is defined by

$$\mathrm{d}_{\mathrm{Was(p)}}(\mu,\mu') := \left(\inf_{\Lambda\in \mathrm{C}(\mu,\mu')} \int_{X\times X} \|\theta-\theta'\|_X^p \mathrm{d}\Lambda(\theta,\theta')\right)^{1/p},$$

where $\mathrm{C}(\mu,\mu') := \{\Lambda'\in \mathrm{Prob}(X^2) : \mu(B)=\Lambda'(B\times X), \mu'(B)=\Lambda'(X\times B), B\in \mathcal{B}X\}$ is the *set of couplings* of $\mu,\mu'\in\mathrm{Prob}_p(X)$. We can link convergence in the Wasserstein(p) distance to weak convergence. Let $(\mu_n)_{n\in\mathbb{N}}\in \mathrm{Prob}_p(X)^{\mathbb{N}}$ be a sequence, and let $\mu\in\mathrm{Prob}_p(X)$ be some other probability measure. Then, according to [80, Theorem 6.9], we have

$$\begin{aligned}(3.2)\qquad &\lim_{n\to\infty} \mathrm{d}_{\mathrm{Was(p)}}(\mu_n,\mu)=0\\ &\Leftrightarrow \left(\lim_{n\to\infty} \mathrm{d}_{\mathrm{Prok}}(\mu_n,\mu)=0 \text{ and } \lim_{n\to\infty}\int \|\theta\|_X^p \mu_n(\mathrm{d}\theta) = \int \|\theta\|_X^p \mu(\mathrm{d}\theta)\right).\end{aligned}$$

Hence, to show Wasserstein(p) well-posedness, we need to show weak well-posedness and stability of the pth posterior moment with respect to changes in the data. Assumptions (A1)–(A4) are not sufficient to show the latter. As in subsection 3.4, we now give the additional assumption (A5) that we need to show Wasserstein(p) well-posedness. Then, we discuss situations in which this assumption is satisfied. We finish this section by showing Wasserstein well-posedness under (A1)–(A5).

ASSUMPTION 3.10. *Consider* (BIP). *Let the following assumption hold:*

(A5) *$g'\in \mathbf{L}^1(X,\mu_{\text{prior}})$ exists such that $\|\theta'\|_X^p \cdot L(y^\dagger|\theta') \leq g'(\theta')$ for μ_{prior}-a.e. $\theta'\in X$ and all $y^\dagger\in Y$.*

Assumption (A5) eventually requires a uniform bound on the pth moment of the posterior measure. This is, in general, not as easily satisfied as (A1)–(A4). However, there is a particular case when we can show that (A1)–(A5) are satisfied rather easily, which is if the likelihood is bounded uniformly by a constant and if the prior has a finite pth moment.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

PROPOSITION 3.11. *We consider* (BIP) *and some* $p \in [1, \infty)$. *Let* (A1) *and* (A4) *hold. Moreover, let some* $c \in (0, \infty)$ *exist, such that*

$$L(y^\dagger|\theta') \leq c \qquad (y^\dagger \in Y; \theta' \in X, \mu_{\text{prior}}\text{-}a.s.),$$

and let $\mu_{\text{prior}} \in \text{Prob}_p(X)$. *Then* (A1)–(A5) *are satisfied.*

We have already mentioned that a uniformly bounded likelihood does not appear to be a very restrictive property. Boundedness of the pth moment of the prior is rather restrictive, however. In practical problems, prior measures very often come from well-known families of probability measures, such as Gaussian, Cauchy, and exponential. For such families we typically know whether certain moments are finite. In this case, it is easy to see, with Proposition 3.11, whether (BIP) satisfies assumption (A5). Hence, (A5) is restrictive but easily verifiable. Next, we state our result on Wasserstein(p) well-posedness.

THEOREM 3.12. *Let* $p \in [1, \infty)$, *and let* (A1)–(A5) *hold for* (BIP). *Then,* (BIP) *is Wasserstein*(p) *well-posed.*

Finally, we note that weak and Wasserstein(p) stability can also hold in degenerate Bayesian inverse problems; see subsection 2.3. Given the argumentation in subsection 3.2, we see that the Bayesian inverse problem discussed in Proposition 2.6 is stable in the weak topology but not in the Hellinger or in the tv sense. Indeed, the Bayesian inverse problem is also stable in the Wasserstein(p) distance for any $p \in [1, \infty)$ but satisfies neither (A1) nor (A4).

COROLLARY 3.13. *We consider the Bayesian inverse problem given in Proposition* 2.6*; i.e., we assume that the posterior measure is given by*

$$\mu^\dagger_{\text{post}} = \delta(\cdot - \mathcal{G}^{-1}(y^\dagger)) \qquad (y^\dagger \in Y),$$

and that $\mathcal{G}^{-1} : Y \to X$ *is continuous. Then this posterior measure is stable in the weak topology. If* X *is additionally a normed space, the posterior is also stable in Wasserstein*(p) *for any* $p \in [1, \infty)$.

4. Well-Posedness in Quasi-semimetrics. The distances we have considered in section 3 ($\text{d}_{\text{Hel}}, \text{d}_{\text{tv}}, \text{d}_{\text{Prok}}, \text{d}_{\text{Was(p)}}$) are all well-defined metrics. In statistics, and especially in information theory, various distance measures are used that are not actually metrics. For instance, they are asymmetric (quasi-metrics), they do not satisfy the triangle inequality (semimetrics), or they do not satisfy either (quasi-semimetrics). Due to their popularity, it is natural to consider stability also in such generalized distance measures.

The *Kullback–Leibler divergence* (KLD) is a popular quasi-semimetric used in information theory and machine learning. In the following, we consider the KLD exemplary as a quasi-semimetric, in which we discuss well-posedness. The KLD is used to describe the *information gain* when going from $\mu \in \text{Prob}(X)$ to another measure $\mu' \in \text{Prob}(X, \mu)$. If defined, it is given by

$$\text{D}_{\text{KL}}(\mu'\|\mu) := \int_X \log\left(\frac{\text{d}\mu'}{\text{d}\mu}\right) \text{d}\mu'.$$

The KLD induces a topology; see [5]. Hence, we can indeed describe continuity in the KLD and, thus, consider the *Kullback–Leibler well-posedness* of Bayesian inverse problems. This concept bridges information theory and Bayesian inverse problems and

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

allows statements about the loss of information in the posterior measure when the data is perturbed. In particular, we define this *loss of information* by the information gain when going from the posterior $\mu^{\ddagger}_{\text{post}}$ with perturbed data $y^{\ddagger}$ to the posterior $\mu^{\dagger}_{\text{post}}$ with unperturbed data $y^{\dagger}$. Hence, the loss of information is equal to $\mathrm{D}_{\mathrm{KL}}(\mu^{\dagger}_{\text{post}} \| \mu^{\ddagger}_{\text{post}})$. A Bayesian inverse problem is Kullback–Leibler well-posed if the posterior measure exists, if it is unique, and if the information loss is continuous with respect to the data.

DEFINITION 4.1 (Kullback–Leibler well-posed). *The problem* (BIP) *is* Kullback–Leibler well-posed *if*

1. $\mu^{\dagger}_{\text{post}} \in \mathrm{Prob}(X, \mu_{\text{prior}})$ *exists* (existence)*;*
2. $\mu^{\dagger}_{\text{post}}$ *is unique in* $\mathrm{Prob}(X, \mu_{\text{prior}})$ (uniqueness)*; and*
3. *for all* $y^{\dagger} \in Y$ *and* $\varepsilon > 0$*, there is* $\delta(\varepsilon) > 0$*, such that*

$$\mathrm{D}_{\mathrm{KL}}(\mu^{\dagger}_{\text{post}} \| \mu^{\ddagger}_{\text{post}}) \leq \varepsilon \quad (y^{\ddagger} \in Y : \|y^{\dagger} - y^{\ddagger}\|_Y \leq \delta(\varepsilon)) \qquad \text{(stability)}.$$

In the setting of Theorem 2.5, (A1)–(A4) are not sufficient to show Kullback–Leibler well-posedness; indeed, the Kullback–Leibler divergence may not even be well defined. We require the following additional assumption on the log-likelihood.

ASSUMPTION 4.2. *Consider* (BIP)*. Let the following assumption hold for* μ_{prior}*-a.e.* $\theta' \in X$ *and every* $y^{\dagger} \in Y$*:*

(A6) *There are a* $\delta > 0$ *and a function* $h(\cdot, y^{\dagger}) \in \mathbf{L}^1(X, \mu^{\dagger}_{\text{post}})$ *such that*

$$|\log L(y^{\ddagger}|\cdot)| \leq h(\cdot, y^{\dagger}) \qquad (y^{\ddagger} \in Y : \|y^{\dagger} - y^{\ddagger}\|_Y \leq \delta).$$

Assumption (A6) is much more restrictive than (A1)–(A4). Indeed, we now require some integrability condition on the forward response operator. The condition may be hard to verify when the posterior measure has heavy tails, when the model is unbounded, or when we are not able to analyze the underlying model.

THEOREM 4.3. *Let* (A1)–(A4) and (A6) *hold for* (BIP)*. Then* (BIP) *is Kullback–Leibler well-posed.*

Remark 4.4. We note that we have allowed the bound in (A6) to depend on $y^{\dagger} \in Y$ and to hold only locally on sets of the form $\{y^{\ddagger} : \|y^{\dagger} - y^{\ddagger}\|_Y \leq \delta\}$, rather than uniformly over Y. In the same way, we can also generalize the given "global" versions of (A3) and (A5) to local versions. This will, for instance, be required in the proof of Corollary 5.3. However, we imagine that in most practical cases the global versions of (A3) and (A5) are not too restrictive. Hence, for the sake of simplicity we prefer those.

5. The Additive Gaussian Noise Case. In practice, the data space is typically finite-dimensional, and a popular modeling assumption for measurement error is additive nondegenerate Gaussian noise. In this case, one can verify assumptions (A1)–(A4)—independently of prior μ_{prior} and forward response operator $\mathcal{G} \in \mathbf{M} = \{f : X \to Y \text{ measurable}\}$. Hence, this very popular setting leads to a well-posed Bayesian inverse problem in the weak topology, the Hellinger distance, and the tv distance. If the prior has a finite pth moment, we additionally obtain Wasserstein(p) well-posedness.

COROLLARY 5.1. *Let* $Y := \mathbb{R}^k$ *and* $\Gamma \in \mathbb{R}^{k \times k}$ *be symmetric positive definite. Let* $\mathcal{G} \in \mathbf{M}$ *be a measurable function. A Bayesian inverse problem with additive nondegenerate Gaussian noise* $\eta \sim \mathrm{N}(0, \Gamma)$ *is given by the following likelihood:*

$$L(y^{\dagger}|\theta) = \det(2\pi\Gamma)^{-1/2} \exp\left(-\frac{1}{2}\|\Gamma^{-1/2}(\mathcal{G}(\theta) - y^{\dagger})\|_Y^2\right).$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Then (BIP) *corresponding to likelihood* L *and*

(a) *any prior* $\mu_{\text{prior}} \in \text{Prob}(X)$ *is whtv well-posed;*

(b) *any prior* $\mu_{\text{prior}} \in \text{Prob}_p(X)$ *is Wasserstein*(p) *well-posed, where* $p \in [1, \infty)$ *and* X *is a normed space.*

Remark 5.2. Let X contain at least two elements. The non-Bayesian inverse problem (IP) corresponding to the additive Gaussian noise setting in Corollary 5.1 is ill-posed. We have shown this in Proposition 2.2. Hence, in the case of Gaussian noise, the Bayesian approach using any prior measure always gives a whtv well-posed Bayesian inverse problem, in contrast to the always ill-posed (IP).

The fact that we can show well-posedness under *any* forward response operator and *any* prior measure in $\text{Prob}(X)$ or $\text{Prob}_p(X)$ has relatively strong implications for practical problems. We now comment on the deterministic discretization of posterior measures, hierarchical models, and Bayesian model selection.

5.1. Deterministic Discretization. Bayesian inverse problems can be discretized with deterministic quadrature rules; such rules are quasi-Monte Carlo [22], sparse grids [70], and Gaussian quadrature. Those are then used to approximate the model evidence and to integrate with respect to the posterior. Deterministic quadrature rules often behave like discrete approximations of the prior measure. If this discrete approximation is a probability measure supported on a finite set, we can apply Corollary 5.1 and show that (BIP) based on the discretized prior is whtv and Wasserstein(p) well-posed for any $p \in [1, \infty)$.

5.2. Hierarchical Prior. *Hierarchical prior measures* are used to construct more complex and flexible prior models. In Bayesian inverse problems, they are discussed in [26, 27, 56]. The basic idea is to employ a prior measure depending on a so-called hyperparameter. This hyperparameter has itself a prior measure, which (typically) leads to a more complex total prior measure. This can be continued recursively down to K layers:

$$\mu_{\text{prior}} = \int_{X_K} \cdots \int_{X_1} \mu^0_{\text{prior}}(\cdot|\theta_1)\mu^1_{\text{prior}}(\mathrm{d}\theta_1|\theta_2) \dots \mu^K_{\text{prior}}(\mathrm{d}\theta_K).$$

Here, $X_1, \dots, X_K$ are measurable subsets of Radon spaces, $X_0 := X$, and

$$\mu^{k-1}_{\text{prior}} : X_k \times \mathcal{B}X_{k-1} \to [0, 1]$$

is a Markov kernel from $(X_k, \mathcal{B}X_k)$ to $(X_{k-1}, \mathcal{B}X_{k-1})$ for $k \in \{1, \dots, K\}$. Note that hierarchical measures are, in a way, the probabilistic version of a deep model in machine learning, such as a deep neural network. In a deep neural network, we also add layers to allow for more flexibility in function approximations.

The likelihood still depends only on θ but not on the deeper layers $\theta_1, \dots, \theta_K$. The (BIP) of determining the posterior measure $\mathbb{P}(\theta \in \cdot|y = y^\dagger)$ of the outer layer is whtv well-posed. This is a direct implication of Corollary 5.1. Moreover, finding the posterior measure of all layers $\mathbb{P}((\theta, \theta_1, \dots, \theta_K) \in \cdot|y = y^\dagger)$ is whtv well-posed, too. This can be seen by extending the parameter space to $X \times X_1 \times \cdots \times X_K$ to all layers (θ_k lives in X_k, $k = 1, \dots, K$) and applying Corollary 5.1 to the extended parameter space.

5.3. Model Selection. In *Bayesian model selection*, not only a model parameter is identified but also the correct model in a collection of possible models. For instance, Lima et al. [61] applied Bayesian model selection to identify the correct model to

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

represent a particular tumor. We briefly comment on a special case of Bayesian model selection. Let $L(\cdot|\theta,\mathcal{G})$ be the likelihood in Corollary 5.1, where we now also note the dependence on the forward response operator $\mathcal{G}$. Moreover, let $\mathbf{M}' \subseteq \mathbf{M}$ be a finite collection of forward response operators from which we want to identify the *correct* one. We now define a prior measure μ'_{prior} on $\mathbf{M}'$ which determines our a priori knowledge about the model choice. The posterior measure of the model selection problem on $(X \times \mathbf{M}', \mathcal{B}X \otimes 2^{\mathbf{M}'})$ is given by

$$\mu^{\dagger,\text{ms}}_{\text{post}} = \mathbb{P}\left((\theta,\mathcal{G}^*) \in \cdot | \mathcal{G}^*(\theta) + \eta = y^\dagger\right),$$

where $\mathcal{G}^* : \Omega \to \mathbf{M}'$ is the random variable representing the model; it satisfies $\mathcal{G}^* \sim \mu'_{\text{prior}}$. The posterior can be computed using a generalization of Bayes' theorem,

$$\mu^{\dagger,\text{ms}}_{\text{post}}(A \times B) = \frac{\sum_{\mathcal{G}\in B} \int_A L(y^\dagger|\theta,\mathcal{G}) \mu'_{\text{prior}}(\{\mathcal{G}\}) \mathrm{d}\mu_{\text{prior}}(\theta)}{\sum_{\mathcal{G}'\in\mathbf{M}'} \int_X L(y^\dagger|\theta',\mathcal{G}') \mu'_{\text{prior}}(\{\mathcal{G}'\}) \mathrm{d}\mu_{\text{prior}}(\theta')} \quad (A \in \mathcal{B}X, B \in 2^{\mathbf{M}'}).$$

This identity is indeed correct: we just apply Theorem 2.5 on the parameter space $X \times \mathbf{M}'$ with prior measure $\mu_{\text{prior}} \otimes \mu'_{\text{prior}}$ and likelihood $L(y^\dagger|\cdot,\cdot) : X \times \mathbf{M}' \to [0,\infty)$. In the setting of Corollary 5.1, (BIP) of the identifying model and parameter is whtv well-posed.

5.4. Generalizations. We have discussed finite-dimensional data and additive nondegenerate Gaussian noise. These results cannot trivially be expanded to the degenerate Gaussian noise case: degenerate Gaussian likelihoods do not satisfy (A1) and can lead to degenerate posterior measures; we have discussed those in subsection 2.3.

The infinite-dimensional data with additive Gaussian noise requires a likelihood definition via the Cameron–Martin theorem. For a discussion of infinite-dimensional data spaces, we refer the reader to [74, Remark 3.8] for compact covariance operators and to [47, section 2.1] specifically for Gaussian white noise generalized random fields. Generalizing the result from [47], we can state the following.

COROLLARY 5.3. *Let $(Y', \langle\cdot,\cdot\rangle_{Y'})$ be a separable Hilbert space, and let $\Gamma : Y' \to Y'$ be a covariance operator; i.e., it is self-adjoint, positive definite, and trace-class. We assume that Y is the Cameron–Martin space of $\mathrm{N}(0,\Gamma) \in \mathrm{Prob}(Y')$, i.e.,*

$$(Y, \langle\cdot,\cdot\rangle_Y) = (\mathrm{img}(\Gamma^{1/2}, Y'), \langle\Gamma^{-1/2}\cdot, \Gamma^{-1/2}\cdot\rangle_{Y'}),$$

where the inverse square-root $\Gamma^{-1/2}$ is well defined. Moreover, let $\mathcal{G} : X \to Y$ be a measurable function. Then the inverse problem

$$\mathcal{G}(\theta^\dagger) + \eta = y^\dagger \qquad (\eta \sim \mathrm{N}(0,\Gamma))$$

can be represented by the likelihood

$$L(y^\dagger|\theta) = \exp\left(\langle\mathcal{G}(\theta), y^\dagger\rangle_Y - \frac{1}{2}\|\mathcal{G}(\theta)\|_Y^2\right).$$

If, in addition, $\mathcal{G} : X \to Y$ is bounded, then (BIP) *corresponding to likelihood L and*

(a) *any prior $\mu_{\text{prior}} \in \mathrm{Prob}(X)$ is whtv well-posed;*
(b) *any prior $\mu_{\text{prior}} \in \mathrm{Prob}_p(X)$ is Wasserstein(p) well-posed, where $p \in [1,\infty)$ and X is a normed space.*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Note that here we require $\mathcal{G}$ to be bounded in X. Hence, while allowing for infinite-dimensional data spaces, we now have conditions on the forward response operator and also on the covariance operator. Thus, this result is not as generally applicable as Corollary 5.1.

Generalizations to non-Gaussian noise models are not as difficult. In the proof of Corollary 5.1, we used only the fact that the pdf of the noise is strictly positive, continuous in its argument, and bounded by a constant. This, however, is also satisfied when the noise is additive and nondegenerate and follows, e.g., the Cauchy distribution, the t-distribution, or the Laplace distribution.

6. Numerical Illustrations.

We illustrate some of the results shown in the previous sections with numerical examples. First, we consider some simple one-dimensional examples complementing the examples we have considered throughout the article. These include Bayesian inverse problems with likelihoods that are discontinuous in parameter or data. Second, we consider an inverse problem that is high-dimensional in terms of data and parameter. The high-dimensional inverse problem is concerned with the reconstruction of an image by Gaussian process regression.

6.1. Discontinuities in the Likelihood.

In previous works, Lipschitz continuity of the log-likelihood in the data and (at least) continuity in the parameter has been assumed; see [74]. In this article, we prove results that do not require continuity in the parameter; however, we still require continuity in the data. We now illustrate these results with simple numerical experiments. Indeed, we show that assumption (A4) is crucial by comparing (BIP) posteriors with likelihoods that are continuous and discontinuous in the data.

Example 6.1 (continuity of $y \mapsto L(y|\cdot)$). We define the data and parameter spaces by $Y := \mathbb{R}$ and $X := [0,1]$. We consider (BIP)s with prior measure $\mu_{\text{prior}} := \text{Unif}(0,1)$ on X and one of the following likelihoods:

(a) $L(y^\dagger|\theta) = (2\pi)^{-1/2} \exp(-\frac{1}{2}\|y^\dagger - \theta\|_Y^2)$;

(b) $L(y^\dagger|\theta) = (2\pi)^{-1/2} \exp(-\frac{1}{2}\|\lfloor y^\dagger \rfloor - \theta\|_Y^2)$.

We solve the inverse problems in Example 6.1 with numerical quadrature. In particular, we compute the model evidence for a $y^\dagger \in \{-5, -4.999, -4.998, \dots, 5\}$ and the Hellinger distances between $\mu^\dagger_{\text{post}}$ and $\mu^\ddagger_{\text{post}}$, where $y^\ddagger = 1$. In Figure 2, we plot the likelihood functions at $\theta = 0$, the logarithms of the posterior densities, and the Hellinger distances. The top row in the figure refers to (a), and the bottom row refers to (b). In the continuous setting (a), we see continuity with respect to $y^\dagger$ in all images. Indeed, (BIP) in (a) fulfills (A1)–(A4). The inverse problem in (b) satisfies (A1)–(A3) but not (A4). Also, we see discontinuities with respect to the data in all images referring to (b). Especially, the image of the Hellinger distances is discontinuous, which leads to the conclusion that this inverse problem is not well-posed. Hence, (A4) is indeed crucial to obtaining well-posedness of a Bayesian inverse problem.

Remark 6.2. A likelihood as in Example 6.1(b) can arise when considering cumulative or categorial data, rather than real-valued continuous data as in (a). Categorial data arises in classification problems.

While continuity in the data is important, we now illustrate that continuity in the forward response operator is not necessary to obtain continuity in the data-to-posterior map. We give an example that can be understood as learning the bias in a single-layer neural network.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

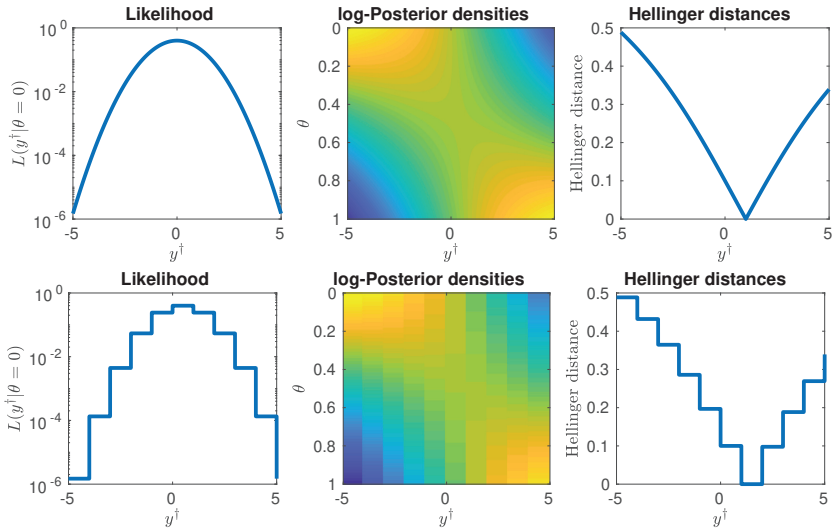

Fig. 2 *Top row: Example* 6.1(a). *Bottom row: Example* 6.1(b). *Left column: Likelihood at* $\theta = 0$. *Center column: Log-posterior densities corresponding to the Bayesian inference problems. The colormaps show a descent in posterior density when going from yellow (high) to dark blue (low). Right column: Hellinger distance between the posterior* $\mu^{\ddagger}_{\text{post}}$ *with* $y^{\ddagger} = 1$ *and posterior* $\mu^{\dagger}_{\text{post}}$ *with* $y^{\dagger}$ *varying between* -5 *and* 5.

Example 6.3 (continuity in $\theta \mapsto L(\cdot|\theta)$). We define the data and parameter spaces by $Y := \mathbb{R}$ and $X := [0, 1]$. Let $w \in [1, \infty]$ be a known weight parameter. We define the forward response operator with weight w by

$$\mathcal{G}_w : X \to Y, \qquad \theta \mapsto \frac{1}{1 + \exp(-w(0.5 - \theta))}.$$

If $w < \infty$, the forward response operator resembles a single-layer neural network with sigmoid activation function evaluated at 0.5. This neural network has known weight w and uncertain bias θ. Moreover, note that in the limiting setting $w = \infty$, the sigmoid function is replaced by the Heaviside function with step at θ, evaluated also at $x = 0.5$:

$$\mathcal{G}_\infty : X \to Y, \qquad \theta \mapsto \begin{cases} 1 & \text{if } 0.5 \geq \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

We consider (BIP) of estimating the true bias $\theta^\dagger$, given an observation $y^\dagger_w := \mathcal{G}_w(\theta^\dagger) + \eta^\dagger$. Here, we consider the noise $\eta^\dagger$ to be a realization of $\eta \sim \mathrm{N}(0, 1^2)$. Moreover, we assume that the parameter $\theta \sim \mu_{\text{prior}} = \mathrm{Unif}(0, 1)$ follows a uniform prior measure.

We solve (BIP)s in Example 6.3 with weights $w = 1, 10, 100, \infty$ again with numerical quadrature for $y^\dagger \in \{-13, -12.99, -12.98, \ldots, 13\}$. We compute the Hellinger distance between $\mu^\dagger_{\text{post}}$ and $\mu^\ddagger_{\text{post}}$, where $y^\ddagger = 0$. In Figure 3, we plot the logarithms of the posterior densities obtained in Example 6.3, along with the Hellinger distances. We observe that all of the posteriors are continuous with respect to the data. These include the posterior that is based on the discontinuous forward response operator

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

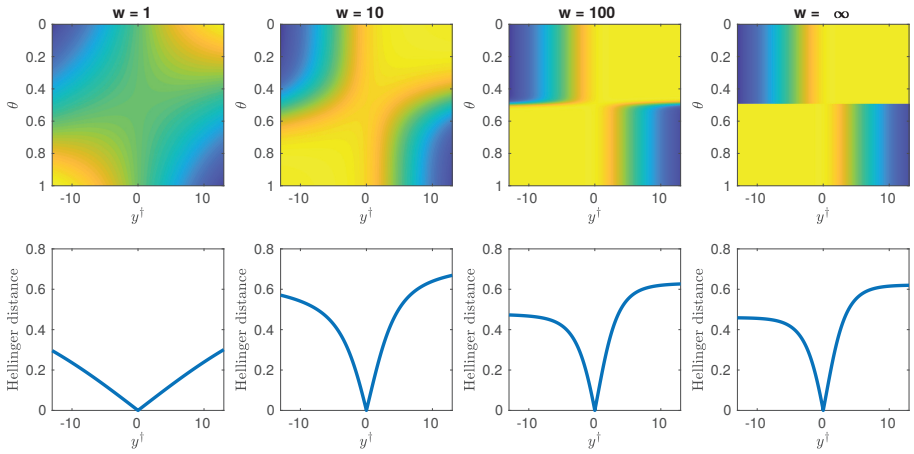

Fig. 3 *From left to right: Example* 6.3 *given* $w = 1, 10, 100, \infty$. *Top row: Log-posterior densities corresponding to the Bayesian inference problems. The colormaps show a descent in posterior density when going from yellow (high) to dark blue (low). Bottom row: Hellinger distance between the posterior* $\mu^{y^\ddagger}_{\text{post}}$ *with* $y^\ddagger = 1$ *and posterior* $\mu^{y^\dagger}_{\text{post}}$ *with* $y^\dagger$ *varying between* -13 *and* 13.

$\mathcal{G}_\infty$. It is discontinuous in the parameter but continuous in the data. (BIP)s considered here satisfy again (A1)–(A4). Hence, these numerical experiments also verify the statement of Lemma 3.7.

Remark 6.4. In *deep learning*, sigmoid functions $\mathcal{G}_w$ $(w < \infty)$ are considered as smooth approximations to the Heaviside function $\mathcal{G}_\infty$, which shall be used as an activation function. The smooth sigmoid functions allow us to train the deep neural network with a gradient-based optimization algorithm. When training the neural network with a Bayesian approach, rather than an optimization approach, we see that we can use Heaviside functions in place of smooth approximations and obtain a well-posed Bayesian inverse problem.

6.2. A High-Dimensional Inverse Problem. We now consider an inverse problem that is high-dimensional in parameter and data spaces. In particular, we observe single, noisy pixels of a grayscale photograph. The inverse problem consists of the reconstruction of the image, for which we use Gaussian process regression. We then perturb the data by adding white noise to the image, and we investigate changes in the posterior as we rescale the noise.

Example 6.5. Let the parameter space $X := \mathbb{R}^{100\times 100}$ contain grayscale images made up of 100×100 pixels. The data space $Y := \mathbb{R}^{25\times 25}$ consists of 25×25 pixels that are observed in a single picture. Returning those 25×25 pixels from a 100×100 pixel image is modeled by the function $\mathcal{G} : X \to Y$. Let $\theta^\dagger \in X$ be a full image. Given

$$y^\dagger = \mathcal{G}(\theta^\dagger) + \eta,$$

we shall recover the full image $\theta^\dagger$. Here, $\eta \sim \mathrm{N}(0, 5^2 I)$ is normally distributed noise, with a noise level of about $5/\max(y) = 2\%$. We assume the following Gaussian prior

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

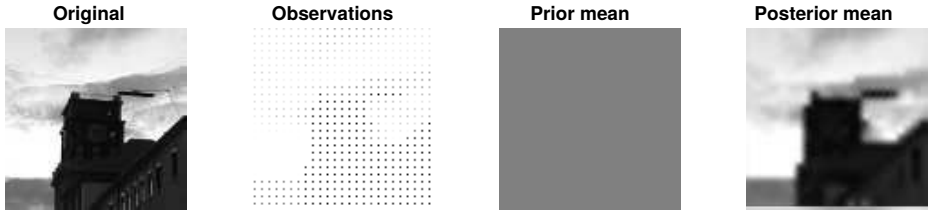

Fig. 4 *Reconstruction of an image with Gaussian process regression. From left to right: Original image, observational data (white parts are unobserved), prior mean, and posterior mean.*

on X:

$$\mu_{\text{prior}} = \mathrm{N}\left(\begin{pmatrix} 128 & \cdots & 128 \\ \vdots & \ddots & \vdots \\ 128 & \cdots & 128 \end{pmatrix}, C_0\right),$$

where $C_0 \in \mathbb{R}^{100\times 4}$ is a covariance tensor assigning the following covariances:

$$\mathrm{Cov}(\theta_{i,j}, \theta_{\ell,k}) = 10000 \cdot \exp\left(-\frac{\sqrt{(i-\ell)^2 + (j-k)^2}}{15}\right).$$

Note that this is essentially an adaptation of an exponential covariance kernel for a Gaussian process in two-dimensional space.

The Bayesian inverse problem in Example 6.5 can be solved analytically, since $\mathcal{G}$ is linear, and the prior and noise are Gaussian. We obtain the posterior measure by Gaussian process regression. In Figure 4, we present the original image, observations, prior mean image, and posterior mean image. The reconstruction is rather coarse, which is not surprising given that we observe only $6.25 \cdot 10^2$ of 10^4 pixels of the image.

We now investigate how the posterior measure changes under marginal changes in the data. To do so, we perturb the image additively with scaled white noise. In particular, we add $\mathrm{N}(0, \sigma^2)$-distributed, independent random variables to each pixel. In Figure 5, we show images and associated observations, where the standard deviation (StD) of the noise is $\sigma \in \{1, 10, 100\}$.

Using Gaussian process regression, we compute the posteriors after perturbing the images with scaled white noise given $\sigma \in \{10^{-17}, 10^{-16}, \ldots, 10^2\}$. Between the original posterior with no perturbation in the data and all others, we compute the Hellinger distance and the relative Frobenius distance between the (matrix-valued) posterior means,

$$\text{relative Frobenius distance } = \frac{\left\| \int \theta \mathrm{d}\mu^{\ddagger}_{\text{post}}(\theta) - \int \theta \mathrm{d}\mu^{\dagger}_{\text{post}}(\theta) \right\|_{\mathrm{F}}}{\left\| \int \theta \mathrm{d}\mu^{\ddagger}_{\text{post}}(\theta) \right\|_{\mathrm{F}}},$$

where $\mu^{\dagger}_{\text{post}}$ (resp., $\mu^{\ddagger}_{\text{post}}$) is the posterior referring to the perturbed data $y^{\dagger}$ (resp., nonperturbed data $y^{\ddagger}$). Since the perturbation is random, we perform this process 20 times and compute the mean over these distances. The standard deviation in these metrics is negligibly small. We plot the results in Figure 6, where we indeed see

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

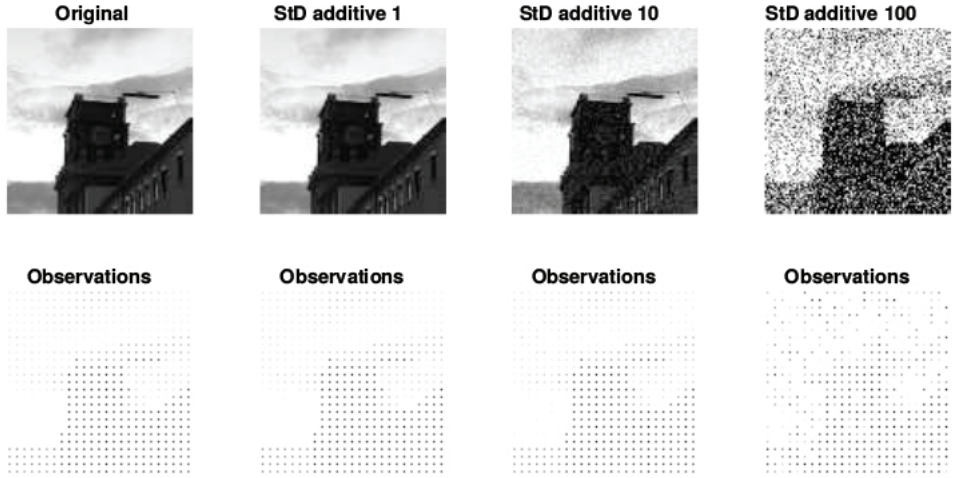

Fig. 5 *Top row: Original image and images perturbed with scaled white noise, given* $\sigma \in \{1, 10, 100\}$. *Bottom row: Observations obtained from the perturbed image.*

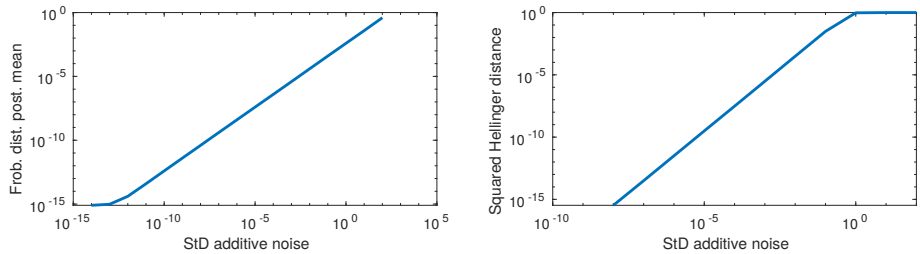

Fig. 6 *Mean relative Frobenius distances and mean squared Hellinger distances computed between the posteriors* $\mu^{\dagger}_{\text{post}}$ *and* $\mu^{\ddagger}_{\text{post}}$, *in which the underlying image was perturbed with white noise that has been scaled by StD* $\sigma = 0, 10^{-17}, 10^{-19}, \ldots, 10^{2}$. *"Mean" refers to the fact that the perturbations are random, and the distances have been computed for* 20 *random perturbations and then averaged. When approaching* $|y^{\ddagger} - y^{\dagger}| \to 0$, *the distances go to* 0. *The left-out* x*-values have distance zero numerically.*

continuity reducing the error StD in the data. In light of Lemma 3.7 and Corollary 5.1, the following is what we expect: First, note that the Bayesian inverse problem falls in the category *additive finite-dimensional Gaussian noise* and is therefore well-posed. Hence, also in this high-dimensional setting, we are able to verify our analytical results concerning well-posedness.

7. Conclusions and Outlook. In this work, we introduce and advocate a new concept of well-posedness of Bayesian inverse problems. We weaken the stability condition by considering continuity instead of Lipschitz continuity of the data-to-posterior map. On the other hand, we make the stability condition somewhat stronger by allowing one to adapt the metric on the space of probability measures to the particular situation. Various notions of well-posedness arise from this discussion; we summarize their relations in Figure 7.

Importantly, we show that given our concept, a huge class of practically relevant

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

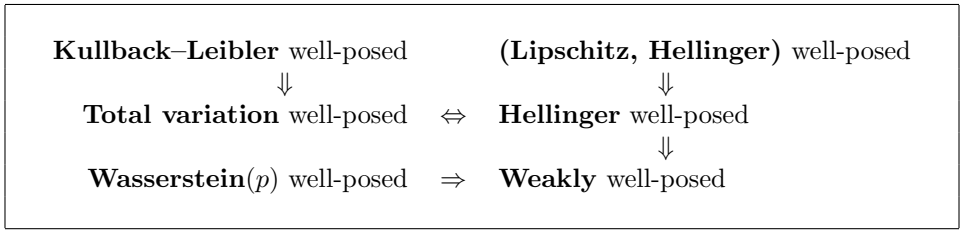

Fig. 7 *Relations between concepts of well-posedness. Here, $A \Rightarrow B$ means that* (BIP) *being A-well-posed implies that it is also B-well-posed.*

Bayesian inverse problems is well-posed or can easily be shown to be well-posed. Hence, we give the following general justification for the Bayesian approach to inverse problems for a huge number of practical situations: here, the Bayesian inverse problem will have a unique solution, which will be robust with respect to marginal changes in the data. Such inverse problems appear, e.g., in the physical and biological sciences, engineering, machine learning, and imaging.

7.1. An Outlook to Recent Developments. Since the publication of the original version of this paper [54], several more developments occurred in the analysis of the well-posedness of Bayesian inverse problems. We briefly summarize some of those works.

Especially in Bayesian image reconstruction, the use of neural-network-based, data-driven prior measures has become popular in recent years. The well-posedness of such approaches has been discussed by [10, 39, 58]. The well-posedness of Bayesian inverse acoustic scattering was studied by [83] in Hellinger and Wasserstein distance, as well as in the Kullback–Leibler divergence. The Hellinger well-posedness of the Bayesian estimation of drift and diffusion coefficient in discretely observed diffusions has been studied by [18].

Lanthaler, Mishra, and Weber [52] show (Lipschitz, Wasserstein)-well-posedness of Bayesian data assimilation even in cases where the underlying forward problem is ill-posed. Similarly, (Lipschitz, Wasserstein)-well-posedness of Bayesian inversion in hyperbolic conservation laws is discussed by [65]. Dolera and Mainini [24] discuss Lipschitz continuity in the Wasserstein distance of general probability kernels. As opposed to previous results that show local Lipschitz continuity, they obtain global Lipschitz continuity. The same authors have also studied the uniform continuity of posterior measures [25] with respect to the total variation distance. The stability of doubly intractable posterior measures, that is, the case where the likelihood itself has an unknown normalizing constant, in terms of Wasserstein(1) and total variation distance, has been studied in [35].

7.2. Directions for Future Research. In the following, we propose some directions for future research.

Degenerate Bayesian Inverse Problems. So far, we have mostly neglected the degenerate Bayesian inverse problems, which we discussed in subsection 2.3. Such problems appear in Bayesian probabilistic numerics [14] and other settings where noise-free data is considered. These may also include the Bayesian formulation of machine learning problems with discrete loss models, such as 0-1-loss, or Bayesian formulations of classification problems; see [6].

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Discrete Data. Inverse problems with discrete data spaces Y appear frequently in applications, such as in computed tomography [72] or when using a charge couple device camera [3]. Usual concepts of well-posedness fail in this situation, as they would need us to discuss continuity of a function on a discrete space. In this case, it may be appropriate to study the difference of posteriors differing in single counts, i.e., where data sets $y^\dagger, y^\ddagger \in Y$ differ exactly by $\|y^\dagger - y^\ddagger\|_1 = 1$.

Approximate Models. In many practical applications, we replace $\mathcal{G}$ by some approximation $\mathcal{G}'$, given, e.g., through a numerical discretization; see, e.g., [44] in the case of differential equations, or an emulator [75]. Asymptotic results about the convergence of posteriors with respect to discretization error are known [63, 73, 74], as is the importance of accurate discretizations in posterior estimations [15, 57]. In addition to the asymptotic results, sharp error estimates with computable constants are necessary for practical applications.

Approximate Posteriors. It is often impossible to find a closed form representation for a posterior, instead computational strategies are needed to obtain an appropriate approximation to the posterior, such as Markov chain Monte Carlo (MCMC). In computationally intensive applications, MCMC may not be an option for the approximation of a posterior measure; instead, inaccurate MCMC methods are used [28, 82] or the posterior is approximated by finding a close representative in a family of probability measures, e.g., in variational Bayes [8], in Bayesian variational autoencoders [50], and in sparse Gaussian process regression [79]. Such approximation techniques need to be analyzed separately with regard to their well-posedness.

Implicit Regularization. As discussed previously, the training of machine learning models can often be understood as an inverse problem. In practice, it is usually too computationally expensive to train a neural network in a Bayesian way. The variational approach (see subsection 2.5) is employed instead. Here, regularization is often done *implicitly*: stochastic optimization techniques with constant stepsizes are employed that have no convergence guarantees in the employed setting [49, 68], but sometimes converge to stationary measures [23, 46, 55]. These stationary measures can be seen as the result of an implicit regularization. Although these stationary measures are often not actual Bayesian posteriors, they are used and interpreted in a similar way [62]. A discussion of the well-posedness of implicitly regularized problems and a better fundamental understanding of implicit regularization in general are necessary and vital for the practical use of machine learning models.

Appendix A. Conditional Probability.

In this appendix, we briefly summarize some results concerning conditional probabilities. Let X, Y be given as in subsection 2.1. Moreover, let $\Omega := X \times Y$, and let $\theta : \Omega \to X, y : \Omega \to Y$ be random variables.

THEOREM A.1. *A Markov kernel $M : Y \times \mathcal{B}X \to [0,1]$ exists, such that*

$$\mathbb{P}(\{\theta \in A\} \cap \{y \in C\}) = \int_C M(y^\dagger, A)\mathbb{P}(y \in \mathrm{d}y^\dagger) \qquad (A \in \mathcal{B}X, C \in \mathcal{B}Y).$$

Moreover, M is $\mathbb{P}(y \in \cdot)$-a.s. unique.

Let $y^\dagger \in Y$. The probability measure $M(y^\dagger, \cdot)$ in Theorem A.1 is the *(regular) conditional probability distribution* of θ given that $y = y^\dagger$. We denote it by $\mathbb{P}(\theta \in \cdot | y = y^\dagger)$. Note that the conditional probability is only unique for a.e. $y^\dagger \in Y$. This definition, as well as Theorem A.1, is nonconstructive. However, if we can represent the joint distribution $\mathbb{P}((\theta, y) \in \cdot)$ by a pdf, we can compute the density of the

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

conditional probability distribution. First, consider the following lemma concerning joint and marginal pdfs.

LEMMA A.2. *Let ν_X and ν_Y be σ-finite measures on $(X, \mathcal{B}X)$ and $(Y, \mathcal{B}Y)$, and let*

$$\mathbb{P}((\theta, y) \in \cdot) \ll \nu_X \otimes \nu_Y \quad \text{with } f := \frac{\mathrm{d}\mathbb{P}((\theta, y) \in \cdot)}{\mathrm{d}\nu_X \otimes \nu_Y} \qquad (\nu_X \times \nu_Y\text{-a.e.}).$$

Then $\mathbb{P}(\theta \in \cdot) \ll \nu_X$, with $\mathrm{d}\mathbb{P}(\theta \in \cdot)/\mathrm{d}\nu_X = \int_X f(\cdot, y^\dagger)\nu_Y(\mathrm{d}y^\dagger)$, ν_X-a.e., and $\mathbb{P}(y \in \cdot) \ll \nu_Y$ with $\mathrm{d}\mathbb{P}(y \in \cdot)/\mathrm{d}\nu_Y = \int_Y f(\theta^\dagger, \cdot)\nu_X(\mathrm{d}\theta^\dagger)$, ν_Y-a.e.

Next, we move on to the construction of the conditional density.

LEMMA A.3. *Let ν_X, ν_Y, and f be given as in Lemma A.2. Then, for $\theta^\dagger \in X$ (ν_X-a.e.) and $y^\dagger \in Y$ (ν_Y-a.e.), we have*

$$\frac{\mathrm{d}\mathbb{P}(\theta \in \cdot | y = y^\dagger)}{\mathrm{d}\nu_X}(\theta^\dagger) = \begin{cases} \frac{f(\theta^\dagger, y^\dagger)}{g(y^\dagger)} & \text{if } g(y^\dagger) > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $g(y^\dagger) := \int_X f(\theta^\ddagger, y^\dagger)\nu_X(\mathrm{d}\theta^\ddagger)$ is the ν_Y-pdf of $\mathbb{P}(y \in \cdot)$.

This result is fundamental to proving Bayes' theorem; see Theorem 2.5.

Appendix B. Proofs. In this appendix, we present rigorous proofs of all of the theorems, propositions, lemmas, and corollaries stated in this article.

Proof of Proposition 2.2. Note that the support of μ_{noise} is all of Y. Hence, the noise $\eta^\dagger$ can be any value in Y, and we need to solve the equation

$$y^\dagger = \mathcal{G}(\theta^\dagger) + \eta^\dagger \tag{B.1}$$

with respect to both $\theta^\dagger \in X$ and $\eta^\dagger \in Y$. Let $\theta' \in X$. Set $\eta' := y^\dagger - \mathcal{G}(\theta')$. Then (θ', η') solves (B.1) and thus the inverse problem (IP). Hence, each element in X implies a solution. Since X contains at least two elements, the solution is not unique, and thus (IP) is ill-posed. ∎

Proof of Theorem 2.5. The following statements hold $\mathbb{P}(y \in \cdot)$-a.s. for $y^\dagger \in Y$.

We first show that $Z(y^\dagger) > 0$. Since we assume that $L(y^\dagger|\cdot)$ is μ_{prior}-a.s. strictly positive, we can write

$$Z(y^\dagger) = \int L(y^\dagger|\theta)\mathrm{d}\mu_{\text{prior}}(\theta) = \int_{\{L(y^\dagger|\cdot)>0\}} L(y^\dagger|\theta)\mathrm{d}\mu_{\text{prior}}(\theta). \tag{B.2}$$

Now let $n \in \mathbb{N}$. As the integrand in (B.2) is positive, Chebyshev's inequality [2, Theorem 2.4.9] implies that

$$n \cdot \int_{\{L(y^\dagger|\cdot)>0\}} L(y^\dagger|\theta)\mathrm{d}\mu_{\text{prior}}(\theta) \geq \mu_{\text{prior}}(L(y^\dagger|\cdot) > n^{-1}). \tag{B.3}$$

We aim to show that the probability on the right-hand side of this equation converges to 1 as $n \to \infty$. Knowing this, we can conclude that the right-hand side is strictly positive for all $n \geq N$ for some $N \in \mathbb{N}$.

Note that measures are continuous with respect to increasing sequences of sets. We define the set

$$B_n := \{L(y^\dagger|\cdot) > n^{-1}\}$$

and observe that $(B_n)_{n=1}^\infty$ is indeed an increasing sequence. Moreover, note that

$$B_\infty = \bigcup_{m=1}^{\infty} B_m = \{L(y^\dagger|\cdot) > 0\}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

and that $\mu_{\text{prior}}(B_\infty) = 1$. Hence, we have

$$\lim_{n\to\infty} \mu_{\text{prior}}(L(y^\dagger|\cdot) > n^{-1}) = \mu_{\text{prior}}(L(y^\dagger|\cdot) > 0) = 1.$$

As mentioned earlier, we now deduce that for some $\varepsilon \in (0,1)$, there is an index $N \in \mathbb{N}$ such that

$$|\mu_{\text{prior}}(L(y^\dagger|\cdot) > n^{-1}) - 1| \leq \varepsilon < 1 \qquad (n \geq N),$$

and thus $\mu_{\text{prior}}(L(y^\dagger|\cdot) > n^{-1}) > 0$ for $n \geq N$. Plugged into (B.3), this gives us $Z(y^\dagger) > 0$. We have also $Z(y^\dagger) < \infty$, since $L(y^\dagger|\cdot) \in \mathbf{L}^1(X, \mu_{\text{prior}})$. Thus, the posterior density (2.3) is well defined. We now apply Bayes' theorem in the formulation of [21, Theorem 3.4] and obtain

$$\frac{\mathrm{d}\mu^\dagger_{\text{post}}}{\mathrm{d}\mu_{\text{prior}}}(\theta') = \frac{L(y^\dagger|\theta')}{Z(y^\dagger)} \qquad (\theta' \in X, \mu_{\text{prior}}\text{-a.s.}).$$

This implies

$$\pi^\dagger_{\text{post}}(\theta') = \frac{\mathrm{d}\mu^\dagger_{\text{post}}}{\mathrm{d}\nu_X}(\theta') = \frac{\mathrm{d}\mu^\dagger_{\text{post}}}{\mathrm{d}\mu_{\text{prior}}}(\theta')\frac{\mathrm{d}\mu_{\text{prior}}}{\mathrm{d}\nu_X}(\theta') = \frac{L(y^\dagger|\theta')\pi_{\text{prior}}(\theta')}{Z(y^\dagger)} \quad (\theta' \in X, \nu_X\text{-a.s.})$$

by application of standard results concerning Radon–Nikodym derivatives. This concludes the proof. ∎

Proof of Proposition 2.6. We test $\mu^\dagger_{\text{post}}$ in Theorem A.1. Let $\theta \sim \mu_{\text{prior}}$ and $y \sim \mu_L(\cdot|\theta)$. Then, $\mathbb{P}(y = \mathcal{G}(\theta)) = 1$. Therefore, for $A \in \mathcal{B}X, C \in \mathcal{B}Y$, we have

$$\begin{aligned}\mathbb{P}(\{\theta \in A\} \cap \{y \in C\}) &= \mathbb{P}(\{y \in \mathcal{G}(A)\} \cap \{y \in C\}) \\ &= \int_C \mathbf{1}_{\mathcal{G}(A)}(y^\dagger)\mathbb{P}(y \in \mathrm{d}y^\dagger) \\ &= \int_C \delta\left(A - \mathcal{G}^{-1}(y^\dagger)\right)\mathbb{P}(y \in \mathrm{d}y^\dagger).\end{aligned}$$

Note that $\mathcal{G}(A) \in \mathcal{B}Y$, since $\mathcal{G}^{-1}$ is continuous. Hence, according to Theorem A.1, we have $\mathbb{P}(\theta \in \cdot|\mathcal{G}(\theta) = y^\dagger) = \delta\left(\cdot - \mathcal{G}^{-1}(y^\dagger)\right)$ for $\mathbb{P}(y \in \cdot)$-a.e. $y^\dagger \in Y$. Moreover, we have $\mathbb{P}(y \in \cdot) = \mu_{\text{prior}}(\mathcal{G} \in \cdot)$. ∎

Proof of Theorem 3.6. Hellinger well-posedness follows from Lemma 3.7. There, we show existence and uniqueness on $\mathrm{P} := \mathrm{Prob}(X, \mu_{\text{prior}})$. According to Theorem 2.5, we again obtain existence and uniqueness of the posterior measure also on $\mathrm{P} := \mathrm{Prob}(X)$, as required for weak and tv well-posedness. By [32], we have

$$\mathrm{d}_{\text{Prok}}(\mu, \mu') \leq \mathrm{d}_{\text{tv}}(\mu, \mu') \leq \sqrt{2}\mathrm{d}_{\text{Hel}}(\mu, \mu') \qquad (\mu, \mu' \in \mathrm{Prob}(X, \mu_{\text{prior}})).$$

Hence, d_{Prok} and d_{tv} are coarser than d_{Hel}. By Proposition 3.9, (BIP) is weakly and tv well-posed. ∎

Proof of Lemma 3.7. Note that existence and uniqueness of the measure $\mu^\dagger_{\text{post}}$ are results of Theorem 2.5 which hold since (A1)–(A2) are satisfied. We proceed as follows: we show that the likelihood is continuous as a function from Y to $\mathbf{L}^1(X, \mu_{\text{prior}})$ and that at the same time $y^\dagger \mapsto Z(y^\dagger)$ is continuous. This implies that $y^\dagger \mapsto L(y^\dagger|\cdot)^{1/2} \in \mathbf{L}^2(X, \mu_{\text{prior}})$ is continuous as well. Then, we collect all of this information and show the continuity in the Hellinger distance, which is the desired result.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

1. We now show continuity in $y^\dagger \in Y$ when integrating $L(y^\dagger|\cdot)$ with respect to μ_{prior}. This is a standard application of *Lebesgue's dominated convergence theorem* (DCT): let $(y_n)_{n=1}^\infty \in Y^{\mathbb{N}}$ be a sequence converging to $y^\dagger$, as $n \to \infty$. Assumption (A4) implies that $\lim_{n\to\infty} L(y_n|\cdot) = L(y^\dagger|\cdot)$ pointwise in X. We obtain by the DCT

$$\lim_{n\to\infty} \int L(y_n|\cdot)\mathrm{d}\mu_{\text{prior}} = \int \lim_{n\to\infty} L(y_n|\cdot)\mathrm{d}\mu_{\text{prior}} = \int L(y^\dagger|\cdot)\mathrm{d}\mu_{\text{prior}},$$

since the sequence $(L(y_n|\cdot))_{n=1}^\infty$ is bounded from above by $g \in \mathbf{L}^1(X, \mu_{\text{prior}})$ and bounded from below by 0; see (A1) and (A3). Hence, the functions

$$Y \ni y^\dagger \mapsto \int L(y^\dagger|\cdot)\mathrm{d}\mu_{\text{prior}} = Z(y^\dagger) \in \mathbb{R}, \qquad Y \ni y^\dagger \mapsto L(y^\dagger|\cdot) \in \mathbf{L}^1(X, \mu_{\text{prior}})$$

are continuous. Moreover, note that Theorem 2.5 implies that $Z(y^\dagger)$ is finite and strictly larger than 0.

2. The continuity in $\mathbf{L}^1(X, \mu_{\text{prior}})$ implies that for every $y^\dagger \in Y$, we have for $\varepsilon_1 > 0$ some $\delta_1(\varepsilon_1) > 0$, such that

$$\|L(y^\dagger|\cdot) - L(y^\ddagger|\cdot)\|_{\mathbf{L}^1(X,\mu_{\text{prior}})} \leq \varepsilon_1 \quad (y^\ddagger \in Y : \|y^\dagger - y^\ddagger\|_Y \leq \delta_1(\varepsilon_1)).$$

Using this, we can show that $y^\dagger \mapsto L(y^\dagger|\cdot)^{1/2}$ is continuous in $\mathbf{L}^2(X, \mu_{\text{prior}})$. Let $y^\dagger \in Y$ and $\varepsilon_1, \delta_1(\varepsilon_1), y^\ddagger$ be chosen as above. We have

$$\begin{aligned}
&\|L(y^\dagger|\cdot)^{1/2} - L(y^\ddagger|\cdot)^{1/2}\|^2_{\mathbf{L}^2(X,\mu_{\text{prior}})} \\
&\qquad = \int \left|L(y^\dagger|\cdot)^{1/2} - L(y^\ddagger|\cdot)^{1/2}\right|^2 \mathrm{d}\mu_{\text{prior}} \\
&\qquad \leq \int \left|L(y^\dagger|\cdot)^{1/2} - L(y^\ddagger|\cdot)^{1/2}\right| \times \left|L(y^\dagger|\cdot)^{1/2} + L(y^\ddagger|\cdot)^{1/2}\right| \mathrm{d}\mu_{\text{prior}} \\
&\qquad = \int \left|L(y^\dagger|\cdot) - L(y^\ddagger|\cdot)\right| \mathrm{d}\mu_{\text{prior}} \leq \varepsilon_1.
\end{aligned}$$

Now, we take the square-root on each side of this inequality. Then, for every $\varepsilon_2 > 0$, choose $\delta_2(\varepsilon_2) := \delta_1(\varepsilon_2^{1/2}) > 0$. Then

$$\|L(y^\dagger|\cdot)^{1/2} - L(y^\ddagger|\cdot)^{1/2}\|_{\mathbf{L}^2(X,\mu_{\text{prior}})} \leq \varepsilon_2 \quad (y^\ddagger \in Y : \|y^\dagger - y^\ddagger\|_Y \leq \delta_2(\varepsilon_2))$$

gives us the desired continuity result.

3. Using the continuity result in item 1 and the composition of continuous functions, we also know that $y^\dagger \mapsto Z(y^\dagger)^{-1/2} \in (0, \infty)$ is continuous. Hence, we have for every $y^\dagger \in Y$ and every $\varepsilon_3 > 0$ a $\delta_3(\varepsilon_3) > 0$ with

$$|Z(y^\dagger)^{-1/2} - Z(y^\ddagger)^{-1/2}| \leq \varepsilon_3 \quad (y^\ddagger \in Y : \|y^\dagger - y^\ddagger\|_Y \leq \delta_3(\varepsilon_3)).$$

Given this and all of the previous results, we now employ a technique that is typically used to prove the continuity of the product of two continuous functions. Let $y^\dagger \in Y$, $\varepsilon_2, \varepsilon_3 > 0$, $\delta_4 = \min\{\delta_2(\varepsilon_2), \delta_3(\varepsilon_3)\}$, and $y^\dagger \in Y$: $\|y^\dagger - y^\ddagger\|_Y \leq \delta_4$. We arrive at

$$\begin{aligned}
\mathrm{d}_{\text{Hel}}(\mu^\dagger_{\text{post}}, \mu^\ddagger_{\text{post}}) &= \|Z(y^\dagger)^{-1/2} L(y^\dagger|\theta)^{1/2} - Z(y^\ddagger)^{-1/2} L(y^\ddagger|\theta)^{1/2}\|_{\mathbf{L}^2(X,\mu_{\text{prior}})} \\
&\leq |Z(y^\ddagger)^{-1/2}| \times \|L(y^\ddagger|\theta)^{1/2} - L(y^\dagger|\theta)^{1/2}\|_{\mathbf{L}^2(X,\mu_{\text{prior}})} \\
&\quad + \|L(y^\dagger|\theta)^{1/2}\|_{\mathbf{L}^2(X,\mu_{\text{prior}})} |Z(y^\ddagger)^{-1/2} - Z(y^\dagger)^{-1/2}| \\
&\leq Z(y^\ddagger)^{-1/2}\varepsilon_2 + Z(y^\dagger)^{1/2}\varepsilon_3 \\
&\leq (Z(y^\dagger)^{-1/2} + \varepsilon_3)\varepsilon_2 + Z(y^\dagger)^{1/2}\varepsilon_3,
\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

where in the last step we have used $|Z(y^\dagger)^{-1/2} - Z(y^\ddagger)^{-1/2}| \leq \varepsilon_3$. We now choose some $\varepsilon_4 > 0$ and set $\delta_4 = \min\{\delta_2(\varepsilon_2'), \delta_3(\varepsilon_3')\}$, where we set

$$\varepsilon_2' := \frac{\varepsilon_4 Z(y^\dagger)^{1/2}}{\varepsilon_4 + 2}, \quad \varepsilon_3' := \frac{\varepsilon_4}{2Z(y^\dagger)^{1/2}}.$$

Then we obtain that $\mathrm{d_{Hel}}(\mu^\dagger_{\mathrm{post}}, \mu^\ddagger_{\mathrm{post}}) \leq \varepsilon_4$ for any $y^\ddagger \in Y$, such that $\|y^\dagger - y^\ddagger\|_Y \leq \delta_4$. This implies the continuity of the posterior measure in Hellinger distance. ∎

Proof of Lemma 3.8. For every $a \in A$ and $\varepsilon > 0$, there is a $\delta(\varepsilon) > 0$, with

$$d_2(f(a), f(a')) \leq \varepsilon \qquad (a' \in A : d_A(a, a') \leq \delta(\varepsilon)).$$

Hence, for the same a, a', ε, and δ, we have

$$d_1(f(a), f(a')) \leq t(d_2(f(a), f(a'))) \leq t(\varepsilon).$$

Since t is continuous in 0, we find for every $\varepsilon' > 0$ some $\delta'(\varepsilon') > 0$, such that $|t(x)| \leq \varepsilon'$ for $x \in [0, \infty) : |x| \leq \delta'(\varepsilon')$. Now, we choose $\delta''(\varepsilon'') := \delta(\delta'(\varepsilon''))$ for every $a \in A$ and $\varepsilon'' > 0$. Then

$$d_1(f(a), f(a')) \leq t(d_2(f(a), f(a'))) \leq t(\delta'(\varepsilon'')) \leq \varepsilon'' \qquad (a' \in A : d_A(a, a') \leq \delta''(\varepsilon'')),$$

which results in continuity in (B, d_1). ∎

Proof of Proposition 3.9. By assumption the Bayesian inverse problem is (P, d_2)-well-posed. Hence, the posterior measure $\mu^\dagger_{\mathrm{post}} \in \mathrm{P}$ exists and is unique. Moreover, the map $Y \ni y^\dagger \to \mu^\dagger_{\mathrm{post}} \in (\mathrm{P}, d_2)$ is continuous. Since d_1 is coarser than d_2, Lemma 3.8 implies that $Y \ni y^\dagger \to \mu^\dagger_{\mathrm{post}} \in (\mathrm{P}, d_1)$ is continuous as well. Hence, the Bayesian inverse problem is (P, d_1)-well-posed. ∎

Proof of Proposition 3.11. We show that (A3) and (A5) hold. Note that (A2) is implied by (A3). We set $g \equiv c$. Then $L \leq g$. Since μ_{prior} is a probability measure, we have

$$\int_X g \mathrm{d}\mu_{\mathrm{prior}} = c\mu_{\mathrm{prior}}(X) = c < \infty.$$

Hence, $g \in \mathbf{L}^1(X, \mu_{\mathrm{prior}})$, which implies that (A3) is satisfied. Next, we define $g'(\theta') := c \cdot \|\theta'\|_X^p$ for $\theta' \in X, \mu_{\mathrm{prior}}$-a.s. By this definition, we have $\|\cdot\|_X^p \cdot L(y^\dagger|\cdot) \leq g'$ for all $y^\dagger \in Y$. Moreover,

$$\int_X g' \mathrm{d}\mu_{\mathrm{prior}} = c \int_X \|\theta\|_X^p \mu_{\mathrm{prior}}(\mathrm{d}\theta) < \infty,$$

since $\mu_{\mathrm{prior}} \in \mathrm{Prob}_p(X)$, and thus $\int_X \|\theta\|_X^p \mu_{\mathrm{prior}}(\mathrm{d}\theta) < \infty$. Hence, $g' \in \mathbf{L}^1(X, \mu_{\mathrm{prior}})$, implying that (A5) holds. ∎

Proof of Theorem 3.12. Let $p \in [1, \infty)$ and $y^\dagger \in Y$. Since (A1)–(A4) hold, we have existence and uniqueness of $\mu^\dagger_{\mathrm{post}} \in \mathrm{Prob}(X)$ by Theorem 3.6. We first show that $\mu^\dagger_{\mathrm{post}} \in \mathrm{Prob}_p(X)$: we have

$$\int_X \|\theta\|_X^p \mu^\dagger_{\mathrm{post}}(\mathrm{d}\theta) = \frac{\int_X L(y^\dagger|\theta)\|\theta\|_X^p \mu_{\mathrm{prior}}(\mathrm{d}\theta)}{\int_X L(y^\dagger|\theta)\mu_{\mathrm{prior}}(\mathrm{d}\theta)} \leq \frac{\int_X g'(\theta)\mu_{\mathrm{prior}}(\mathrm{d}\theta)}{\int_X L(y^\dagger|\theta)\mu_{\mathrm{prior}}(\mathrm{d}\theta)} < \infty,$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

where the left-hand side is bounded by Theorem 2.5 (denominator) and by (A5) (numerator). Hence, the posterior measure exists in $\mathrm{Prob}_p(X)$. Since $\mathrm{Prob}_p(X) \subseteq \mathrm{Prob}(X)$, the posterior measure is also unique in $\mathrm{Prob}_p(X)$. Hence, existence and uniqueness of the posterior are satisfied.

Now, we move on to stability. As in the proof of Lemma 3.7, the map $Y \ni y^\dagger \mapsto Z(y^\dagger) \in (0,\infty)$ is continuous. By the DCT and (A5), the map

$$Y \ni y^\dagger \mapsto \int_X L(y^\dagger|\theta)\|\theta\|_X^p \mu_{\mathrm{prior}}(\mathrm{d}\theta) \in [0,\infty)$$

is continuous as well. Therefore,

$$\begin{aligned}\int_X \|\theta\|_X^p \mu_{\mathrm{post}}^\dagger(\mathrm{d}\theta) &= \frac{\int_X L(y^\dagger|\theta)\|\theta\|_X^p \mu_{\mathrm{prior}}(\mathrm{d}\theta)}{\int_X L(y^\dagger|\theta)\mu_{\mathrm{prior}}(\mathrm{d}\theta)} \\ &\to \frac{\int_X L(y^\ddagger|\theta)\|\theta\|_X^p \mu_{\mathrm{prior}}(\mathrm{d}\theta)}{\int_X L(y^\ddagger|\theta)\mu_{\mathrm{prior}}(\mathrm{d}\theta)} = \int_X \|\theta\|_X^p \mu_{\mathrm{post}}^\ddagger(\mathrm{d}\theta)\end{aligned}$$

as $y^\dagger \to y^\ddagger$. Hence, we have stability of the posterior measure in the pth moment. Additionally, we have weak well-posedness due to Theorem 3.6, and thus stability in the $\mathrm{d}_{\mathrm{Prok}}$. By (3.2), we have stability in $\mathrm{d}_{\mathrm{Was(p)}}$.

Therefore, we also have Wasserstein(p) well-posedness of (BIP). ∎

Proof of Corollary 3.13. According to Proposition 2.6, the posterior measure $\mu_{\mathrm{post}}^\dagger$ is well defined and unique. Let $f : X \to \mathbb{R}$ be bounded and continuous. Then

$$\lim_{y^\ddagger \to y^\dagger} \int f \mathrm{d}\mu_{\mathrm{post}}^\ddagger = \lim_{y^\ddagger \to y^\dagger} f \circ \mathcal{G}^{-1}(y^\ddagger) = f \circ \mathcal{G}^{-1}(y^\dagger) = \int f \mathrm{d}\mu_{\mathrm{post}}^\dagger, \tag{B.4}$$

since $f \circ \mathcal{G}^{-1}$ is continuous. Therefore, $Y \ni y^\dagger \mapsto \mu_{\mathrm{post}}^\dagger \in (\mathrm{Prob}(X), \mathrm{d}_{\mathrm{Prok}})$ is continuous. Thus, we have weak well-posedness. If now X is a normed space and $p \in [1,\infty)$, the mapping $\|\cdot\|_X^p : X \to \mathbb{R}$ is continuous. Note that when setting $f := \|\cdot\|_X^p$ in (B.4), the equation still holds. Thus, we have stability in the pth moment and therefore also Wasserstein(p) well-posedness according to (3.2). ∎

Proof of Theorem 4.3. First, note that (A1)–(A4) imply the existence and uniqueness of the posterior measure, as well as the continuity of $y^\dagger \mapsto Z(y^\dagger)$. Let $y^\dagger \in Y$ and $y^\ddagger \in Y$, with $\|y^\dagger - y^\ddagger\|_Y \le \delta$. $\delta > 0$ is chosen as in (A6). We have

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\mu_{\mathrm{post}}^\dagger \| \mu_{\mathrm{post}}^\ddagger) &= \int \log\left(\frac{\mathrm{d}\mu_{\mathrm{post}}^\dagger}{\mathrm{d}\mu_{\mathrm{post}}^\ddagger}\right) \mathrm{d}\mu_{\mathrm{post}}^\dagger \\ &= \int \log L(y^\dagger|\cdot) - \log L(y^\ddagger|\cdot) \mathrm{d}\mu_{\mathrm{post}}^\dagger + \left(\log Z(y^\ddagger) - \log Z(y^\dagger)\right),\end{aligned}$$

where the right-hand side of this equation is well defined since $Z(y^\dagger), Z(y^\ddagger) \in (0,\infty)$ by Lemma 3.7 and since (A6) holds. Moreover, the continuity in the model evidence implies that $\left(\log Z(y^\ddagger) - \log Z(y^\dagger)\right) \to 0$, as $y^\ddagger \to y^\dagger$. Also, note that $\log L(\cdot\,|\theta')$ is continuous by (A4), which implies

$$\lim_{y^\ddagger \to y^\dagger} \int_X \log L(y^\dagger|\cdot) - \log L(y^\ddagger|\cdot) \mathrm{d}\mu_{\mathrm{post}}^\dagger = \int_X \lim_{y^\ddagger \to y^\dagger} \log L(y^\dagger|\cdot) - \log L(y^\ddagger|\cdot) \mathrm{d}\mu_{\mathrm{post}}^\dagger = 0,$$

where we applied the DCT with $2h(\cdot, y^\dagger)$ as a dominating function. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Proof of Corollary 5.1. We check (A1)–(A4).

(A1) By definition, the likelihood is a strictly positive pdf for any $\theta' \in X$.

(A2)–(A3) The likelihood is bounded above uniformly by $g \equiv \det(2\pi\Gamma)^{-1/2}$ which is integrable with respect to any probability measure on $(X, \mathcal{B}X)$.

(A4) The likelihood is continuous in $y^\dagger$ for any $\theta' \in X$. ∎

Proof of Corollary 5.3. 1. The function L is indeed a correct likelihood, i.e., $y^\dagger \mapsto L(y^\dagger|\theta')$ is a pdf for μ_{prior}-a.e. $\theta' \in X$. We refer the reader to the discussions of the Cameron–Martin theorem in [9, section 2.4] and [76, section 2.7]. Moreover, we again mention [74, Remark 3.8] and [47, section 2.1], which have discussed the modeling in this case. Hence, (A1) is true.

2. Now, we check (A2)–(A4). Note that (A4) is true by assumption. (A2) holds since $\mathcal{G}$ is bounded. (A3) cannot be shown easily. However, we can replace it by a local version of this assumption; see Remark 4.4. Indeed, to show continuity of $\mu^\dagger_{\text{post}}$ in $y^\dagger \in Y$, we only need to satisfy (A3) in $\overline{B}(y^\dagger, \delta) := \{\| \cdot - y^\dagger\|_Y \le \delta\} := \{y^\ddagger : \|y^\ddagger - y^\dagger\|_Y \le \delta\}$ for $\delta > 0$. If we show this for any $y^\dagger \in Y$ and some $\delta > 0$, we obtain stability as well. Note that we have used this idea to show Kullback–Leibler well-posedness in Theorem 4.3.

3. Let $y^\dagger \in Y$ be arbitrary. Let c be chosen such that $\|\mathcal{G}(\theta')\|_Y < c$, which exists since $\mathcal{G}$ is bounded. Let $y^\ddagger \in \overline{B}(y^\dagger, \delta)$. By the Cauchy–Schwarz and triangle inequalities, we have

$$\begin{aligned} L(y^\ddagger|\theta') &= \exp\left(\langle \mathcal{G}(\theta'), y^\ddagger\rangle_Y - \frac{1}{2}\|\mathcal{G}(\theta')\|_Y^2\right) \le \exp\left(|\langle \mathcal{G}(\theta'), y^\ddagger\rangle_Y|\right) \\ &\le \exp\left(\|\mathcal{G}(\theta')\|_Y \|y^\ddagger\|_Y\right) \le \exp(c\|y^\ddagger\|_Y) \\ &= \exp(c\|y^\ddagger - y^\dagger + y^\dagger\|_Y) \le \exp(c\|y^\ddagger - y^\dagger\|_Y + \|y^\dagger\|_Y) \\ &\le \exp(c \cdot (\delta + \|y^\dagger\|_Y)) =: c' \end{aligned}$$

for μ_{prior}-a.e. $\theta' \in X$. Now, we choose $g : X \to \mathbb{R}$ to be $g \equiv c'$. Let now $\mu_{\text{prior}} \in \text{Prob}(X)$. Then, $g \in \mathbf{L}^1(X, \mu_{\text{prior}})$ and $L(y^\ddagger|\theta') \le g(\theta')$ for μ_{prior}-a.e. $\theta' \in X$. Since $y^\dagger$ is chosen arbitrarily, we obtain stability in the weak topology, the Hellinger distance, and the tv distance. Hence, we have whtv well-posedness, and thus we have shown (a).

4. Let $p \in [1, \infty)$. To show Wasserstein(p) well-posedness, we can again use a local argument on the data space. Hence, we can satisfy (A5) locally on the data space. This, on the other hand, is implied by a local version of Proposition 3.11. Hence, we obtain stability in the Wasserstein(p) distance if $\mu_{\text{prior}} \in \text{Prob}_p(X)$ and if for all $y^\dagger \in Y$, we have some $\delta > 0$ and $c' > 0$ such that

$$L(y^\ddagger|\theta') \le c' \qquad (\|y^\ddagger - y^\dagger\|_Y \le \delta; \mu_{\text{prior}}\text{-a.e. } \theta' \in X).$$

This, however, is what we have shown already in item 3. Thus, we have shown (b). ∎

Proof of Theorem A.1. The theorem above holds if Ω, X, Y are Radon spaces; see [59, Theorem 3.1]. Y is a Radon space by definition. X and Ω can be extended to Radon spaces X' and $\Omega' = X' \times Y$, where $\mathbb{P}(\theta \in X' \backslash X) = 0 = \mathbb{P}(\Omega' \backslash \Omega)$. Moreover, we set $M(y, X' \backslash X) = 0$ $(y \in Y)$. ∎

Proof of Lemma A.2. Let $A \in \mathcal{B}X$. Note that

$$\mathbb{P}(\theta \in A) = \mathbb{P}((\theta, y) \in A \times Y) = \int_{A\times Y} f \mathrm{d}(\nu_X \otimes \nu_Y) = \int_A \int_Y f(\theta^\dagger, y^\dagger)\nu_Y(\mathrm{d}y^\dagger)\nu_X(\mathrm{d}\theta^\dagger),$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

where the last equality holds due to Tonelli. Hence, indeed,

$$\frac{\mathrm{d}\mathbb{P}(\theta \in \cdot)}{\mathrm{d}\nu_X} = \int_X f(\cdot, y^\dagger)\nu_Y(\mathrm{d}y^\dagger) \quad (\nu_X\text{-a.e.}).$$

The statement about the ν_Y-pdf of y can be shown by exchanging y and θ, and X and Y. ∎

Proof of Lemma A.3. For a derivation in the case $X := Y := \mathbb{R}$, see [2, Example 5.3.2 (b)]. The proof in our more general setting is analogous. ∎

Acknowledgments. The author thanks several contributors for their highly appreciated support towards the original version of this article [54]: Elisabeth Ullmann made insightful and valuable comments that contributed to this work, as did Björn Sprungk through illuminating discussions with the author. Tim J. Sullivan detected an error in an older version of the manuscript. Florian Beiser and Brendan Keith proofread the article. Lukas Latz helped the author to interpret Hadamard's original work. Both anonymous reviewers made valuable comments which helped to improve the presentation.

REFERENCES

[1] S. Agapiou, A. M. Stuart, and Y.-X. Zhang, *Bayesian posterior contraction rates for linear severely ill-posed inverse problems*, J. Inverse Ill-Posed Probl., 22 (2014), pp. 297–321, https://doi.org/10.1515/jip-2012-0071. (Cited on p. 840)

[2] R. B. Ash and C. Doléans-Dade, *Probability & Measure Theory*, 2nd ed., Harcourt/Academic Press, Burlington, MA, 2000. (Cited on pp. 834, 856, 862)

[3] J. M. Bardsley, *Stopping rules for a nonnegatively constrained iterative method for ill-posed Poisson imaging problems*, BIT, 48 (2008), pp. 651–664, https://doi.org/10.1007/s10543-008-0196-6. (Cited on p. 855)

[4] T. Bayes, *An essay towards solving a Problem in the Doctrine of Chances*, Philos. Trans. Roy. Soc. London, 53 (1763), pp. 370–418. (Cited on p. 835)

[5] R. V. Belavkin, *Asymmetric topologies on statistical manifolds*, in Geometric Science of Information, F. Nielsen and F. Barbaresco, eds., Springer International, Cham, 2015, pp. 203–210, https://doi.org/10.1007/978-3-319-25040-3_23. (Cited on p. 845)

[6] A. L. Bertozzi, X. Luo, A. M. Stuart, and K. C. Zygalakis, *Uncertainty quantification in graph-based classification of high dimensional data*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 568–595, https://doi.org/10.1137/17M1134214. (Cited on p. 854)

[7] P. Billingsley, *Probability and Measure*, 3rd ed., Wiley, 2008. (Cited on p. 834)

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006, https://link.springer.com/book/9780387310732. (Cited on p. 855)

[9] V. I. Bogachev, *Gaussian Measures*, American Mathematical Society, Providence, RI, 1998, https://doi.org/10.1090/surv/062. (Cited on p. 861)

[10] P. Bohra, T. A. Pham, J. Dong, and M. Unser, *Bayesian inversion for nonlinear imaging models using deep generative priors*, IEEE Trans. Comput., 8 (2022), pp. 1237–1249, https://doi.org/10.1109/TCI.2023.3236155. (Cited on p. 854)

[11] K. Bredies and D. Lorenz, *Mathematical Image Processing*, Birkhäuser, Cham, 2018, https://doi.org/10.1007/978-3-030-01458-2. (Cited on p. 838)

[12] G. Chavent, *Nonlinear Least Squares for Inverse Problems: Theoretical Foundations and Step-by-Step Guide for Applications*, Springer, Dordrecht, 2010, https://doi.org/10.1007/978-90-481-2785-6. (Cited on p. 838)

[13] J. Cheng and B. Hofmann, *Regularization methods for ill-posed problems*, in Handbook of Mathematical Methods in Imaging, O. Scherzer, ed., Springer, New York, 2015, pp. 91–123, https://doi.org/10.1007/978-1-4939-0790-8_3. (Cited on p. 838)

[14] J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami, *Bayesian probabilistic numerical methods*, SIAM Rev., 61 (2019), pp. 756–789, https://doi.org/10.1137/17M1139357. (Cited on pp. 837, 854)

[15] J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami, *Probabilistic numerical methods for PDE-constrained Bayesian inverse problems*, AIP Conf. Proc., 1853 (2017), art. 060001, https://doi.org/10.1063/1.4985359. (Cited on p. 855)

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

[16] D. R. Cox, *Principles of Statistical Inference*, Cambridge University Press, 2006. (Cited on p. 836)
[17] R. T. Cox, *Probability, Frequency and Reasonable Expectation*, Amer. J. Phys., 14 (1946), pp. 1–13, https://doi.org/10.1119/1.1990764. (Cited on p. 835)
[18] J. C. Croix, M. Dashti, and I. Z. Kiss, *Nonparametric Bayesian Inference of Discretely Observed Diffusions*, preprint, https://arxiv.org/abs/2004.04636, 2000. (Cited on p. 854)
[19] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss, *MAP estimators and their consistency in Bayesian nonparametric inverse problems*, Inverse Problems, 29 (2013), art. 095017, https://doi.org/10.1088/0266-5611/29/9/095017. (Cited on p. 839)
[20] M. Dashti and A. M. Stuart, *Uncertainty quantification and weak approximation of an elliptic inverse problem*, SIAM J. Numer. Anal., 49 (2011), pp. 2524–2542, https://doi.org/10.1137/100814664. (Cited on pp. 833, 841)
[21] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, in Handbook of Uncertainty Quantification, R. Ghanem, D. Higdon, and H. Owhadi, eds., Springer, Cham, 2017, pp. 311–428, https://doi.org/10.1007/978-3-319-12385-1_7. (Cited on pp. 833, 836, 840, 841, 857)
[22] J. Dick, R. N. Gantner, Q. R. Le Gia, and C. Schwab, *Higher order quasi-Monte Carlo integration for Bayesian PDE inversion*, Comput. Math. Appl., 77 (2019), pp. 144–172, https://doi.org/10.1016/j.camwa.2018.09.019. (Cited on p. 847)
[23] A. Dieuleveut, A. Durmus, and F. Bach, *Bridging the gap between constant step size stochastic gradient descent and Markov chains*, Ann. Statist., 48 (2020), pp. 1348–1382, https://doi.org/10.1214/19-AOS1850. (Cited on p. 855)
[24] E. Dolera and E. Mainini, *Lipschitz Continuity of Probability Kernels in the Optimal Transport Framework*, preprint, https://arxiv.org/abs/2010.08380, 2020. (Cited on p. 854)
[25] E. Dolera and E. Mainini, *On uniform continuity of posterior distributions*, Statist. Probab. Lett., 157 (2020), art. 108627, https://doi.org/10.1016/j.spl.2019.108627. (Cited on p. 854)
[26] M. M. Dunlop, M. A. Girolami, A. M. Stuart, and A. L. Teckentrup, *How deep are deep Gaussian processes?*, J. Mach. Learn. Res., 19 (2018), pp. 1–46, http://jmlr.org/papers/v19/18-015.html. (Cited on p. 847)
[27] M. M. Dunlop, M. A. Iglesias, and A. M. Stuart, *Hierarchical Bayesian level set inversion*, Statist. Comput., 27 (2017), pp. 1555–1584, https://doi.org/10.1007/s11222-016-9704-8. (Cited on p. 847)
[28] A. Durmus and E. Moulines, *High-dimensional Bayesian inference via the unadjusted Langevin algorithm*, Bernoulli, 25 (2013), pp. 2854–2882, https://doi.org/10.3150/18-BEJ1073. (Cited on p. 855)
[29] S. Engel, D. Hafemeyer, C. Münch, and D. Schaden, *An application of sparse measure valued Bayesian inversion to acoustic sound source identification*, Inverse Problems, 35 (2019), art. 075005, https://doi.org/10.1088/1361-6420/ab1497. (Cited on p. 833)
[30] O. G. Ernst, B. Sprungk, and H.-J. Starkloff, *Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 823–851, https://doi.org/10.1137/140981319. (Cited on p. 833)
[31] A. L. Gibbs, *Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration*, Stoch. Models, 20 (2004), pp. 473–492, https://doi.org/10.1081/STM-200033117. (Cited on p. 843)
[32] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, Internat. Statist. Rev., 70 (2002), pp. 419–435, https://doi.org/10.1111/j.1751-5823.2002.tb00178.x. (Cited on pp. 843, 844, 857)
[33] E. Giné and R. Nickl, *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Cambridge University Press, 2016, https://doi.org/10.1017/CBO9781107337862. (Cited on p. 836)
[34] C. Groetsch, *Linear inverse problems*, in Handbook of Mathematical Methods in Imaging, O. Scherzer, ed., Springer, New York, 2015, pp. 3–46, https://doi.org/10.1007/978-1-4939-0790-8_1. (Cited on p. 838)
[35] M. Habeck, D. Rudolf, and B. Sprungk, *Stability of doubly-intractable distributions*, Electron. Commun. Probab., 25 (2020), art. 62, https://doi.org/10.1214/20-ECP341. (Cited on p. 854)
[36] J. Hadamard, *Sur les problèmes aux dérivés partielles et leur signification physique*, Princeton Univ. Bull., 13 (1902), pp. 49–52. (Cited on p. 832)
[37] T. Helin and M. Burger, *Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems*, Inverse Problems, 31 (2015), art. 085009, https://doi.org/10.1088/0266-5611/31/8/085009. (Cited on p. 839)
[38] E. Hellinger, *Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen*, J. Reine Angew. Math., 136 (1909), pp. 210–271, http://eudml.org/doc/

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

149313. (Cited on p. 838)
[39] M. Holden, M. Pereyra, and K. C. Zygalakis, *Bayesian imaging with data-driven priors encoded by neural networks*, SIAM J. Imaging Sci., 15 (2022), pp. 892–924, https://doi.org/10.1137/21M1406313. (Cited on p. 854)
[40] B. Hosseini, *Well-posed Bayesian inverse problems with infinitely divisible and heavy-tailed prior measures*, SIAM/ASA J. Uncertain. Quantif., 5 (2017), pp. 1024–1060, https://doi.org/10.1137/16M1096372. (Cited on p. 833)
[41] B. Hosseini and N. Nigam, *Well-posed Bayesian inverse problems: Priors with exponential tails*, SIAM/ASA J. Uncertain. Quantif., 5 (2017), pp. 436–465, https://doi.org/10.1137/16M1076824. (Cited on p. 833)
[42] M. A. Iglesias, K. Lin, and A. M. Stuart, *Well-posed Bayesian geometric inverse problems arising in subsurface flow*, Inverse Problems, 30 (2014), art. 114001, https://doi.org/10.1088/0266-5611/30/11/114001. (Cited on pp. 833, 841)
[43] M. A. Iglesias, Y. Lu, and A. M. Stuart, *A Bayesian level set method for geometric inverse problems*, Interfaces Free Bound., 18 (2016), pp. 181–217, https://doi.org/10.4171/IFB/362. (Cited on p. 833)
[44] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, 2008, https://doi.org/10.1017/CBO9780511995569. (Cited on p. 855)
[45] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, 2003, https://doi.org/10.1017/CBO9780511790423. (Cited on p. 835)
[46] K. Jin, J. Latz, C. Liu, and C.-B. Schönlieb, *Continuous-Time Stochastic Gradient Descent for Continuous Data*, preprint, https://arxiv.org/abs/2112.03754, 2021. (Cited on p. 855)
[47] C. Kahle, K. F. Lam, J. Latz, and E. Ullmann, *Bayesian parameter identification in Cahn–Hilliard models for biological growth*, SIAM/ASA J. Uncertain. Quantif., 7 (2019), pp. 526–552, https://doi.org/10.1137/18M1210034. (Cited on pp. 833, 848, 861)
[48] T. Karvonen and C. J. Oates, *Maximum likelihood estimation in Gaussian process regression is ill-posed*, J. Mach. Learn. Res., 24 (2023), pp. 1–47, https://www.jmlr.org/papers/v24/22-1153.html. (Cited on p. 838)
[49] D. P. Kingma and J. Ba, *ADAM: A method for stochastic optimisation*, in Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), 2015, https://arxiv.org/abs/1412.6980. (Cited on p. 855)
[50] D. P. Kingma and M. Welling, *Auto-encoding variational Bayes*, in Proceedings of the International Conference on Learning Representations, 2014, https://doi.org/10.48550/arXiv.1312.6114. (Cited on p. 855)
[51] A. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer-Verlag, 1933. (Cited on p. 835)
[52] S. Lanthaler, S. Mishra, and F. Weber, *On Bayesian data assimilation for PDEs with ill-posed forward problems*, Inverse Problems, 38 (2022), art. 085012, https://doi.org/10.1088/1361-6420/ac7acd. (Cited on p. 854)
[53] M. Lassas and S. Siltanen, *Can one use the total variation prior for edge-preserving Bayesian inversion*, Inverse Problems, 20 (2004), art. 1537, https://doi.org/10.1088/0266-5611/20/5/013. (Cited on p. 839)
[54] J. Latz, *On the well-posedness of Bayesian inverse problems*, SIAM/ASA J. Uncertain. Quantif., 8 (2020), pp. 451–482, https://doi.org/10.1137/19M1247176. (Cited on pp. 831, 834, 854, 862)
[55] J. Latz, *Analysis of stochastic gradient descent in continuous time*, Statist. Comput., 31 (2021), art. 39, https://doi.org/10.1007/s11222-021-10016-8. (Cited on p. 855)
[56] J. Latz, M. Eisenberger, and E. Ullmann, *Fast sampling of parameterised Gaussian random fields*, Comput. Methods Appl. Mech. Engrg., 348 (2019), pp. 978–1012, https://doi.org/10.1016/j.cma.2019.02.003. (Cited on pp. 833, 841, 847)
[57] J. Latz, I. Papaioannou, and E. Ullmann, *Multilevel Sequential^2 Monte Carlo for Bayesian Inverse Problems*, J. Comput. Phys., 368 (2018), pp. 154–178, https://doi.org/10.1016/j.jcp.2018.04.014. (Cited on p. 855)
[58] R. Laumont, V. De Bortoli, A. Almansa, J. Delon, A. Durmus, and M. Pereyra, *Bayesian imaging using Plug & Play priors: When Langevin meets Tweedie*, SIAM J. Imaging Sci., 15 (2022), pp. 701–737, https://doi.org/10.1137/21M1406349. (Cited on p. 854)
[59] D. Leao, Jr., M. Fragoso, and P. Ruffino, *Regular conditional probability, disintegration of probability and Radon spaces*, Proyecciones, 23 (2004), pp. 15–29, https://doi.org/10.4067/S0716-09172004000100002. (Cited on p. 861)
[60] H. C. Lie and T. J. Sullivan, *Equivalence of weak and strong modes of measures on topological vector spaces*, Inverse Problems, 34 (2018), art. 115013, https://doi.org/10.1088/1361-6420/aadef2. (Cited on p. 839)

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[61] E. A. B. F. Lima, J. T. Oden, D. A. Hormuth, T. E. Yankeelov, and R. C. Almeida, *Selection, calibration, and validation of models of tumor growth*, Math. Models Methods Appl. Sci., 26 (2016), pp. 2341–2368, https://doi.org/10.1142/S021820251650055X. (Cited on p. 847)

[62] S. Mandt, M. Hoffman, and D. Blei, *Stochastic gradient descent as approximate Bayesian inference*, J. Mach. Learn. Res., 18 (2017), pp. 1–35, https://www.jmlr.org/papers/volume18/17-214/17-214.pdf. (Cited on p. 855)

[63] Y. Marzouk and D. Xiu, *A stochastic collocation approach to Bayesian inference in inverse problems*, Commun. Comput. Phys., 6 (2009), pp. 826–847, https://global-sci.org/intro/article_detail/cicp/7708.html. (Cited on p. 855)

[64] P. McCullagh, *What is a statistical model? With comments and a rejoinder by the author*, Ann. Statist., 30 (2002), pp. 1225–1310, https://doi.org/10.1214/aos/1035844977. (Cited on p. 836)

[65] S. Mishra, D. Ochsner, A. M. Ruf, and F. Weber, *Well-Posedness of Bayesian Inverse Problems for Hyperbolic Conservation Laws*, preprint, https://arxiv.org/abs/2107.09701, 2021. (Cited on p. 854)

[66] Y. V. Prokhorov, *Convergence of random processes and limit theorems in probability theory*, Theory Probab. Appl., 1 (1956), pp. 157–214, https://doi.org/10.1137/1101016. (Cited on p. 842)

[67] G. R. Richter, *An inverse problem for the steady state diffusion equation*, SIAM J. Appl. Math., 41 (1981), pp. 210–221, https://doi.org/10.1137/0141016. (Cited on p. 841)

[68] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407, https://doi.org/10.1214/aoms/1177729586. (Cited on p. 855)

[69] D. Rudolf and N. Schweizer, *Perturbation theory for Markov chains via Wasserstein distance*, Bernoulli, 24 (2018), pp. 2610–2639, https://doi.org/10.3150/17-BEJ938. (Cited on p. 843)

[70] C. Schwab and A. M. Stuart, *Sparse deterministic approximation of Bayesian inverse problems*, Inverse Problems, 28 (2012), art. 045003, https://doi.org/10.1088/0266-5611/28/4/045003. (Cited on p. 847)

[71] W. Schwarz, *No interpretation of probability*, Erkenntnis, 83 (2018), pp. 1195–1212, https://doi.org/10.1007/s10670-017-9936-9. (Cited on p. 835)

[72] P. M. Shikhaliev, T. Zu, S. Molloi, *Photon counting computing tomography: Concept and initial results*, Med. Phys., 32 (2005), pp. 427–436, https://doi.org/10.1118/1.1854779. (Cited on p. 855)

[73] B. Sprungk, *On the local Lipschitz stability of Bayesian inverse problems*, Inverse Probl., 36 (2020), art. 055015, https://doi.org/10.1088/1361-6420/ab6f43. (Cited on pp. 834, 855)

[74] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559, https://doi.org/10.1017/S0962492910000061. (Cited on pp. 833, 836, 838, 841, 848, 849, 855, 861)

[75] A. M. Stuart and A. L. Teckentrup, *Posterior consistency for Gaussian process approximations of Bayesian posterior distributions*, Math. Comp., 87 (2017), pp. 721–753, https://doi.org/10.1090/mcom/3244. (Cited on p. 855)

[76] T. J. Sullivan, *Introduction to Uncertainty Quantification*, Springer, Cham, 2015, https://doi.org/10.1007/978-3-319-23395-6. (Cited on p. 861)

[77] T. J. Sullivan, *Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors*, Inverse Probl. Imaging, 11 (2017), pp. 857–874, https://doi.org/10.3934/ipi.2017040. (Cited on p. 833)

[78] T. J. Sullivan, *Well-posedness of Bayesian inverse problems in quasi-Banach spaces with stable priors*, PAMM. Proc. Appl. Math. Mech., 17 (2017), pp. 871–874, https://doi.org/10.1002/pamm.201710402. (Cited on p. 833)

[79] M. Titsias, *Variational learning of inducing variables in sparse Gaussian processes*, in Proceedings of the 12th Conference on Artificial Intelligence and Statistics, PMLR, 2009, pp. 567–574, https://proceedings.mlr.press/v5/titsias09a.html. (Cited on p. 855)

[80] C. Villani, *Optimal Transport: Old and New*, Springer, Berlin, 2009, https://doi.org/10.1007/978-3-540-71050-9. (Cited on p. 844)

[81] R. von Mises, *Probability, Statistics, and Truth*, Macmillan, 1939. (Cited on p. 835)

[82] M. Welling and Y. W. Teh, *Bayesian learning via stochastic gradient Langevin dynamics*, in Proceedings of the 28th International Conference on Machine Learning, Omnipress, 2011, pp. 681–688, https://dl.acm.org/doi/10.5555/3104482.3104568. (Cited on p. 855)

[83] X.-M. Yang, Z.-L. Deng, and A. Qian, *Bayesian approach for limited-aperture inverse acoustic scattering with total variation prior*, Appl. Anal., (2022), pp. 1–16, https://doi.org/10.1080/00036811.2022.2116014. (Cited on p. 854)

Downloaded 11/06/23 to 137.189.49.34 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.