2152–5080/17/$35.00

STABILITY AND CONVERGENCE OF SOLUTIONS TO STOCHASTIC INVERSE PROBLEMS USING APPROXIMATE PROBABILITY DENSITIES

Troy Butler,[1] Rylan Spence,[2] Tim Wildey,[3] & Tian Yu Yen[4,]*

[1]*Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO 80202, troy.butler@ucdenver.edu*

[2]*Oden Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, TX 78712, rylan.spence@utexas.edu*

[3]*Computational Mathematics Department, Center for Computing Research, Sandia National Labs, Albuquerque, NM 87185, tmwilde@sandia.gov*

[4]*Scientific Machine Learning Department, Center for Computing Research, Sandia National Labs, Albuquerque, NM 87185, tyen@sandia.gov*

*Address all correspondence to: Tian Yu Yen, Scientific Machine Learning Department, Sandia National Labs, 1450 Innovation Pkwy SE, Albuquerque, NM 87185, E-mail: tyen@sandia.gov

Data-consistent inversion is designed to solve a class of stochastic inverse problems where the solution is a pullback of a probability measure specified on the outputs of a quantities of interest (QoI) Map. This work presents stability and convergence results for the case where finite QoI data result in an approximation of the solution as a density. Given their popularity in the literature, separate results are proven for three different approaches to measuring discrepancies between probability measures: f-divergences, integral probability metrics, and L^p metrics. In the context of integral probability metrics, we also introduce a pullback probability metric that is well-suited for data-consistent inversion. This fills a theoretical gap in the convergence and stability results for data-consistent inversion that have mostly focused on convergence of solutions associated with approximate maps. Numerical results are included to illustrate key theoretical results with intuitive and reproducible test problems that include a demonstration of convergence in the measure-theoretic "almost" sense.

KEY WORDS: *uncertainty quantification, inverse problem, push-forward measure, pullback measure, data consistent, convergence, f-divergence, integral probability metrics*

1. INTRODUCTION

Uncertainty Quantification (UQ) has become a critically important field of study due to the increasing reliance on physics-based computational models to make data-informed and data-consistent decisions. UQ problems are generally categorized as being either forward or inverse problems depending on the direction that uncertainty is considered to propagate. The solutions to these UQ problems are often represented as probability densities, on either model input or output spaces, and often require some form of approximation, which introduces error. The focus of this paper is on the impact of such approximation error on the solutions to a specific class of stochastic inverse problems involving aleatoric (i.e., irreducible) uncertainties where the inferential target is a distribution on model inputs. Specifically, we consider the solution to this class of problems as being defined by a pullback of an observed probability measure associated with specified Quantities of Interest (QoI) defined on the space of model outputs.

Data-consistent inversion (DCI) provides a measure-theoretic framework for solving this class of stochastic inverse problems [5,11,12]. In DCI, the solution has what is referred to as the data-consistency property in that its push-forward through the QoI map matches the observed probability measure. In [12], a density-based solution is derived via the Disintegration Theorem [18]. This particular representation of the solution has seen the most development, analysis, and application in recent years, e.g., see [9,16,38–40,44,47,48,55]. It is worth noting that the density form of the solution perhaps first appeared in [36] where it was derived through heuristic arguments based on logarithmic pooling and referred to as "Bayesian melding." Fundamental distinctions in assumptions, form, and properties of the solution from the typical Bayesian framework led to a distinction of the terminology used in the DCI framework in [13] (which is a follow-up to [12]). In [13] and many of the works that chronologically follow it, an initial and predicted density are used to describe the initial quantification of uncertainties on parameters and QoI, respectively, independent of any observed data. The observed density describes the quantification of uncertainty for the observed QoI data. An update to the initial density is then obtained via the product of the initial density with the ratio of observed to predicted densities evaluated on the outputs of the QoI map. The updated density serves as an exact solution to the aleatoric stochastic inverse problem. In practice, when the observed or predicted densities are not known exactly, they are estimated from finite samples, which results in an approximation to the updated density. This work provides the stability and convergence analysis of approximate updated densities associated with a wide range of common density estimation techniques that we may utilize for estimating the observed or predicted densities.

To situate DCI within the UQ literature, we contrast this framework with the typical Bayesian framework that begins with an initial assumption of epistemic (i.e., reducible) uncertainty in data and parameters. For instance, a common assumption in a Bayesian setting is that noisy data are observed for a single instance of a system associated with true, but unknown, parameter values, e.g., see [4,17,19,22,24,30,31]. The solution to the resulting inverse problem within the Bayesian framework is known as a posterior, which is a conditional density defined by the product of a prior density on parameters and a data-likelihood function that is usually constructed from the differences in simulated and observed QoI data. The posterior does not satisfy the data-consistency property but instead is interpreted as defining the relative likelihoods that any particular estimate for the parameters could have produced all of the observed (noisy) data. Subsequently, the posterior is typically utilized to produce a parameter estimate such as the maximum a posteriori (MAP) estimate, e.g., see [1,10,35]. Convergence analysis in Bayesian frameworks is typically focused on the particular point estimate produced and its associated uncertainty as quantified by the posterior covariance. Such analysis will often make use of the Bernstein-von Mises theorem [46], which guarantees that the resulting uncertainty in a parameter estimate, such as the MAP point, is reduced as more data are incorporated. This is fundamentally distinct from the type of stability and convergence analysis we consider in the DCI framework where the goal is to estimate the entire updated density. We refer the interested reader to either the review paper [5] or Section 7 of [12] for more thorough discussions and examples that compare and contrast these frameworks designed to solve different types of problems.

Prior studies such as [12] provide the theory of existence, uniqueness (up to the choice of initial), and stability of the updated density with respect to perturbations in the various densities. However, that work considered stability only with respect to the L^1-norm, i.e., the total-variation metric. Subsequent studies investigated the stability and convergence of updated densities in L^p (for $1 \leq p \leq \infty$) when the QoI map is subject to epistemic errors due to an approximation of the map using discretized computational models or surrogate representations, e.g., see [13,16]. In this work, we fill a theoretical gap in the DCI literature concerning stability and convergence of solutions when predicted or observed densities are approximated from finite data. While non-parametric kernel density estimation (KDE) is perhaps the most common approach to approximate densities in the DCI literature, there is a growing body of literature on other data-driven approaches for density estimation which utilize different metrics or divergences to analyze convergence rates and optimize approximations, e.g., see [21,25,27,42,43,45,53]. Building upon this growing body of literature, we consider three different classes of stability and convergence results. First, we prove the stability of the updated density with respect to f-divergences. Next, we prove that convergence of the approximate observed or predicted densities in an integral probability metric implies convergence of the updated density in a novel pullback integral probability metric. Finally, we show that the convergence of approximate observed or predicted densities in the L^p metric implies convergence of the updated density in the L^p metric.

The remainder of this paper is organized as follows. In Section 2, we summarize the density-based DCI approach and current L^1-based stability theory. We also provide some direct generalizations of the assumptions and theory that set the stage for the more novel results provided in subsequent sections. In Section 3, we consider the class of f-divergences, prove a general result regarding the f-divergence between the initial and updated distributions, and prove stability of the updated density in the f-divergence with respect to approximations of the observed and predicted densities. Then, in Section 4, we consider the class of integral probability metrics (IPM), prove stability in the IPM with respect to appoximations of the observed and predicted densities, and introduce a novel pullback IPM. In Section 5, we prove stability and convergence of the updated density in the L^p-metric. Numerical demonstrations of key theoretical results are provided in Section 6 and concluding remarks are found in Section 7.

2. DATA-CONSISTENT INVERSION

Let $\Lambda \subset \mathbb{R}^n$ denote the parameters of interest in a particular simulation model and $(\Lambda, \mathcal{B}_\Lambda, \mu_\Lambda)$ the associated measure space using the Borel σ-algebra $\mathcal{B}_\Lambda$ and Lebesgue measure μ_Λ. We denote the quantities of interest (QoI) map as $Q : \Lambda \to \mathcal{D} \subset \mathbb{R}^d$ where $(\mathcal{D}, \mathcal{B}_\mathcal{D}, \mu_\mathcal{D})$ is the measure space of possible observed data with $\mathcal{D} := Q(\Lambda)$ denoting the image of Λ.

A standard assumption is that Q is measurable so that $Q^{-1}(E) \in \mathcal{B}_\Lambda$ for all $E \in \mathcal{B}_\mathcal{D}$ where Q^{-1} denotes the pre-image map, which is a common notation used in measure theory. We emphasize that Q is not assumed to be invertible since in general d and n need not be equal.

The stochastic inverse problem is now defined as follows.

Definition 2.1.

Given an observed probability measure, $\mathbb{P}_{\text{obs}}$ on $(\mathcal{D}, \mathcal{B}_\mathcal{D})$, the stochastic inverse problem is to find a probability measure, $\mathbb{P}_\Lambda$ on $(\Lambda, \mathcal{B}_\Lambda)$, that is data-consistent in the sense that

$$\mathbb{P}_\Lambda(Q^{-1}(E)) = \mathbb{P}_{\text{obs}}(E), \tag{1}$$

for all events $E \in \mathcal{B}_{\mathcal{D}}$.

Assuming $\mathbb{P}_\Lambda$ and $\mathbb{P}_{\text{obs}}$ are absolutely continuous with respect to μ_Λ and $\mu_{\mathcal{D}}$, respectively (i.e., assuming probability densities, π_Λ and π_{obs}, exist), then the stochastic inverse problem above is equivalent to finding a density π_Λ such that

$$\mathbb{P}_\Lambda(Q^{-1}(E)) = \int_{Q^{-1}(E)} \pi_\Lambda(\lambda)\, \mu_\Lambda = \int_E \pi_{\text{obs}}(q)\, \mu_{\mathcal{D}} = \mathbb{P}_{\text{obs}}(E),\ \forall E \in \mathcal{B}_{\mathcal{D}}. \tag{2}$$

In either case, the solution to the stochastic inverse problem is a *pullback* probabiltiy measure. This is equivalent to saying that the observed probability measure should be the push-forward of the solution to the stochastic inverse problem. When both Q is one-to-one (implying $d = n$) and the Jacobian of Q exists, then the stochastic inverse problem has a unique solution that can be determined, in theory, by the classical change of variables formula:

$$\pi_\Lambda(\lambda) = \pi_{\text{obs}}(Q(\lambda))\ |J_Q|$$

where $|J_Q|$ is the determinant of the Jacobian of $Q(\lambda)$. One of the main challenges of solving the stochastic inverse problem is that QoI maps are typically ill-posed, i.e., $Q^{-1}(q)$ is not unique for a given $q \in \mathcal{D}$. This is often true even if $d = n$ due to nonlinearities in the map. In [12], a measure-theoretic framework based on the disintegration theorem [20] is developed and analyzed for constructing a density-based solution, which we summarize below.

2.1 Density-based solutions

Since the stochastic inverse problem is in general ill-posed due to the potential existence of many pullback measures, the framework of [12] utilizes an initial density, denoted by π_{init}, defined on $(\pi_\Lambda, \mathcal{B}_\Lambda)$ to regularize the space of solutions. The push-forward of π_{init} through the QoI map defines the predicted density, π_{pred}, i.e.,

$$\mathbb{P}_{\text{pred}}(E) := \int_E \pi_{\text{pred}}(q)\mu_{\mathcal{D}} = \int_{Q^{-1}(E)} \pi_{\text{init}}(\lambda)\mu_\Lambda = \mathbb{P}_{\text{init}}(Q^{-1}(E)) \tag{3}$$

for every event $E \in \mathcal{B}_{\mathcal{D}}$. If π_{init} leads to a predicted density π_{pred} that is equal to π_{obs} almost everywhere, then π_{init} is itself a data-consistent solution to the stochastic inverse problem. However, making such an a priori choice for π_{init} is unrealistic. Instead, we utilize the predicted density to construct an update to the initial density that is data-consistent.

Definition 2.2.

Given both an observed density, π_{obs}, and an initial density, π_{init}, with corresponding predicted density, π_{pred}, the updated density is defined as

$$\pi_{\text{up}}(\lambda) := \pi_{\text{init}}(\lambda) r(\lambda), \quad \text{where} \quad r(\lambda) = \frac{\pi_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))}. \tag{4}$$

The proofs of existence, uniqueness, and stability of the updated density is a consequence of the disintegration theorem [8], which rewrites integrals in a convenient form for the analysis of pullback measures, where for any $A \in \mathcal{B}_\Lambda$,

$$\mathbb{P}_{\text{up}}(A) := \int_A \pi_{\text{up}}(\lambda)\, \mu_\Lambda = \int_{\mathcal{D}} \left(\int_{A \cap Q^{-1}(q)} \pi_{\text{init}}(\lambda) \frac{\pi_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))} d\mu_{\Lambda,q} \right) \mu_{\mathcal{D}}. \tag{5}$$

Here, $\mu_{\Lambda,q}$ denotes the disintegration of the Lebesgue measure μ_Λ along the set $\Lambda \cap Q^{-1}(q) := \{\lambda \in \Lambda : Q(\lambda) = q\}$. To see that $\mathbb{P}_{\text{up}}$ defines a consistent solution, we need to show that $\mathbb{P}_{\text{up}}(Q^{-1}(B)) = \mathbb{P}_{\text{obs}}(B)$ for every $B \in \mathcal{B}_{\mathcal{D}}$. To show this, we first observe that for each $q \in \mathcal{D}$, $Q(\lambda) = q$ in the inner integral since $\lambda \in \Lambda \cap Q^{-1}(q)$. This implies that the observed and predicted densities can be factored out of the inner integral with $Q(\lambda)$ replaced by q. The inner integral subsequently integrates to $\pi_{\text{pred}}(q)$, which cancels the denominator of the factored out ratio and results in an integral over B of $\pi_{\text{obs}}(q)$. It immediately follows that $\mathbb{P}_{\text{up}}(Q^{-1}(B)) = \mathbb{P}_{\text{obs}}(B)$. See [12] for more details. The proofs in this present work make extensive use of the disintegration of measures.

An important theoretical detail is that a predictability assumption is required for the updated density to be a data-consistent solution to the stochastic inverse problem. In its weakest form, the assumption is that π_{obs} is absolutely continuous with respect to π_{pred}. However, in practice, we often assume a stronger form, which we state below.

Assumption 1.

There exists a constant $C > 0$ such that

$$\pi_{\text{obs}}(q) \le C \pi_{\text{pred}}(q), \quad \text{for a.e. } q \in \mathcal{D}.$$

Intuitively, this assumption requires that the support of the predicted density contains the support of the observed density. At a more practical level, this form of the predictability assumption guarantees that standard random sampling schemes can be utilized (see [12] for more details). This form also guarantees that any observed values with positive likelihood are likely to be predicted by the choice of QoI map and initial density. Loosely speaking, we must be able to predict the observed data with push-forward samples from the initial density through the QoI map. Note that this means that the constant C is implicitly dependent on the initial density and the QoI map: a different choice of π_{init} or Q leads to a different predictability constant.

Assumption 1 is straightforward to verify in practice by first noting that

$$\mathbb{E}_{\text{init}}(r(\lambda)) = \int_\Lambda r(\lambda)\pi_{\text{init}}(\lambda)\mu_\Lambda = \int_\Lambda \pi_{\text{up}}(\lambda)\mu_\Lambda = 1.$$

In other words, if the predictability assumption holds, then the updated density is in fact a density implying its integral is equal to one. If samples are generated from the initial probability measure, then this expectation can be approximated as follows

$$\mathbb{E}(r(\lambda)) = \int_\Lambda r(\lambda)\pi_{\text{init}}(\lambda)\mu_\Lambda \approx \frac{1}{N}\sum_{i=1}^{N} r(\lambda_i). \tag{6}$$

Thus, comparing the sample average of the updated ratios to one provides a convenient computational diagnostic to verify the predictability assumption is satisfied. While outside the scope of the current work, if the predictability assumption is violated, recent methods on formulating the problem within a variational framework and utilizing gradient flows to shift the support of the initial density may prove useful, e.g., see [33].

We conclude this particular subsection with the following definition of a conditional density on $\Lambda \cap Q^{-1}(q)$ for a given $q \in \mathcal{D}$ that is useful in the proofs of this paper.

Definition 2.3.

For $q \in \mathcal{D}$ with $\pi_{\text{pred}}(q) > 0$ we define

$$\pi_{\text{init}\,|q}(\lambda) := \frac{\pi_{\text{init}}(\lambda)}{\pi_{\text{pred}}(q)} \tag{7}$$

to be the initial probability density conditioned on q.

We note that $\pi_{\text{init}\,|q}$ is a valid probability density over the set $\Lambda \cap Q^{-1}(q)$ due to the fact that π_{pred} is the push forward of π_{init}, i.e.,

$$\pi_{\text{pred}}(q) = \int_{\Lambda \cap Q^{-1}(q)} \pi_{\text{init}}(\lambda)\, d\mu_{\Lambda,q} \quad \forall q \in \mathcal{D}. \tag{8}$$

It is also worth noting that in (4), π_{up} involves a re-weighting of π_{init} by the ratio, $r(\lambda)$, of the observed and predicted densities that are both evaluated at $Q(\lambda)$. As a consequence, if λ is restricted to a parameter set where $Q(\lambda) = q$ for some fixed $q \in \mathcal{D}$, then π_{up} is simply a re-scaling of π_{init}. This implies that π_{up} and π_{init} have exactly the same conditional densities when conditioned on q. In other words, $r(\lambda)$ serves to update the initial density only in those directions informed by the QoI data.

2.2 Stability and Convergence: Total Variation (TV) Metric

It is often the case that the observed and predicted densities, and therefore the updated density, are numerically approximated using a finite number of samples from these distributions. Prior work (e.g., see [12,15]) on assessing the impact of these approximations utilized the total variation (TV) metric, which is sometimes referred to as the L^1-metric on the space of probability measures defined on a common measure space that are all absolutely continuous with respect to the same dominating measure.

Definition 2.4.
Let $\mathbb{P}^A$ and $\mathbb{P}^B$ represent probability measures on the measure space $(X, \mathcal{B}_X, \mu_X)$ that admit Radon-Nikodym derivatives (with respect to μ_X) $\pi^A(x)$ and $\pi^B(x)$, respectively. Then, the total variation (TV) metric is given by

$$d_{TV}(\mathbb{P}^A, \mathbb{P}^B) := \int_X |\pi^A(x) - \pi^B(x)| d\mu_X. \tag{9}$$

Throughout this paper, we assume that either the observed or predicted densities are approximated in some manner. The following theorems, paraphrased from [12], involve the stability of the updated density with respect to perturbations in the observed or predicted densities. An important note is that the TV metrics involving the observed or predicted densities are computed over $(\mathcal{D}, \mathcal{B}_{\mathcal{D}})$ while the TV metrics involving the updated densities are computed over $(\Lambda, \mathcal{B}_\Lambda)$.

Theorem 1 (Predicted Stability in TV)**.** *For fixed measures $\mathbb{P}_{init}$ and $\mathbb{P}_{obs}$ with corresponding densities π_{init} and*

π_{obs}, respectively, let $\widetilde{\pi}_{pred}$ denote an approximation to π_{pred} such that

$$\pi_{obs}(q) \leq C\widetilde{\pi}_{pred}(q), \quad \text{for a.e. } q \in \mathcal{D},$$

for some constant $C > 0$, and let $\widetilde{\mathbb{P}}_{up}$ denote the associated updated measure obtained from this approximation. Then,

$$d_{TV}(\mathbb{P}_{up}, \widetilde{\mathbb{P}}_{up}) \leq C d_{TV}(\mathbb{P}_{pred}, \widetilde{\mathbb{P}}_{pred}).$$

Proof. See the proof of Theorem 5.1 in [12]. □

Theorem 1 justifies the approximation of the predicted density using finite samples drawn from the initial density and propagated through the QoI map. Specifically, it guarantees that such errors will go to zero as long as $\widetilde{\pi}_{\text{pred}}$ converges to π_{pred} in the limit of infinite samples. In other words, $\widetilde{\pi}_{\text{up}} \to \pi_{\text{up}}$ in $L^1(\Lambda)$ as $\widetilde{\pi}_{\text{pred}} \to \pi_{\text{pred}}$ in $L^1(\mathcal{D})$. Note that the convergences occur in different spaces.

Theorem 2 (Observed Stability in TV)**.** *For fixed measures $\mathbb{P}_{init}$ and $\mathbb{P}_{pred}$ with corresponding densities π_{init} and π_{pred}, respectively, let $\widetilde{\mathbb{P}}_{obs}$ denote an approximation to $\mathbb{P}_{obs}$ such that*

$$\widetilde{\pi}_{obs}(q) \leq C\pi_{pred}(q), \quad \text{for a.e. } q \in \mathcal{D},$$

for some constant $C > 0$, and let $\widetilde{\mathbb{P}}_{up}$ denote the associated updated measure obtained from this approximation. Then,

$$d_{TV}\left(\mathbb{P}_{up}, \widetilde{\mathbb{P}}_{up}\right) = d_{TV}\left(\mathbb{P}_{obs}, \widetilde{\mathbb{P}}_{obs}\right).$$

Proof. See the proof of Theorem 4.1 in [12]. □

Theorem 2 states that the approximation error in the observed density is exactly the approximation error of the corresponding approximation of the updated density. It immediately follows that $\widetilde{\pi}_{\text{up}} \to \pi_{\text{up}}$ in $L^1(\Lambda)$ as $\widetilde{\pi}_{\text{obs}} \to \pi_{\text{obs}}$ in $L^1(\mathcal{D})$.

2.3 Direct Generalization of TV Results

The objective of the remainder of this paper is to generalize the stability and convergence results mentioned above to other divergences and metrics that quantify the discrepancy between two probability measures.

In several cases, the TV metric is noted as a special case. Before we proceed to these generalizations, we note that Theorems 1 and 2 involve comparing a single approximation of the updated density to the exact updated density. Here, we generalize these results to compare two separate updated probability densities associated with two distinct approximations to either the observed or predicted densities. We make use of this generalization to analyze stability and convergence with f-divergences and integral probability metrics in Sections 3 and 4. It also serves as the basis for constructing some of the numerical examples in Section 6. First, we require a generalization of the predictability assumption.

Assumption 2.

There exists a constant $C > 0$ such that:

1. Given arbitrary observed densities π_{obs}^A and π_{obs}^B

$$\pi_{\text{obs}}^A(q) \leq C\pi_{\text{pred}}(q), \quad \text{and} \quad \pi_{\text{obs}}^B(q) \leq C\pi_{\text{pred}}(q) \quad \text{for a.e. } q \in \mathcal{D}.$$

2. Given arbitrary predicted densities π_{pred}^A and π_{pred}^B

$$\pi_{\text{obs}}(q) \leq C\pi_{\text{pred}}^A(q), \quad \text{and} \quad \pi_{\text{obs}}(q) \leq C\pi_{\text{pred}}^B, \quad \text{for a.e. } q \in \mathcal{D}.$$

Note that when two approximations to an observed or predicted density are considered, Assumption 2 provides conditions that guarantee that each of the associated updated density approximations also exist. In many of the theorems below, Assumptions 2.1 and 2.2 are also utilized to provide useful bounds for various terms in the proofs. In cases involving Assumption 2.2, we often require an additional assumption that one of the approximated predicted densities can be scaled to bound the exact predicted density (and without loss of generality, we make this assumption for π_{pred}^A). This allows us to handle technical complications that arise in the proofs related to the predicted density appearing in the denominator of the updated density.

Theorem 3. *For fixed measures $\mathbb{P}_{init}$ and $\mathbb{P}_{obs}$ with corresponding densities π_{init} and π_{obs} respectively, let $\mathbb{P}_{pred}^A$ and $\mathbb{P}_{pred}^B$ denote arbitrary predicted measures which satisfy Assumption 2.2 with associated updated measures $\mathbb{P}_{up}^A$ and*

$\mathbb{P}_{up}^B$. *Additionally, assume there exists another constant* $C_1 > 0$ *such that*

$$\pi_{pred}(q) \leq C_1 \pi_{pred}^A(q), \quad \text{for a.e. } q \in \mathcal{D}.$$

Then, there exists a constant $C_2 > 0$ *such that*

$$d_{TV}(\mathbb{P}_{up}^A, \mathbb{P}_{up}^B) \leq C_2 d_{TV}(\mathbb{P}_{pred}^A, \mathbb{P}_{pred}^B).$$

Proof. See APPENDIX A.1. □

As mentioned previously, for two different predicted densities, we require the additional assumption that the true predicted density is absolutely continuous with respect to π_{pred}^A. This assumption is not necessary for the case of two different observed densities.

Theorem 4. *For fixed measures* $\mathbb{P}_{init}$ *and* $\mathbb{P}_{pred}$ *with corresponding densities* π_{init} *and* π_{pred} *respectively, let* $\mathbb{P}_{obs}^A$ *and* $\mathbb{P}_{obs}^B$ *denote arbitrary observed measures which satisfy Assumption 2.1 with associated updated measures* $\mathbb{P}_{up}^A$ *and* $\mathbb{P}_{up}^B$. *Then,*

$$d_{TV}\left(\mathbb{P}_{up}^A, \mathbb{P}_{up}^B\right) = d_{TV}\left(\mathbb{P}_{obs}^A, \mathbb{P}_{obs}^B\right).$$

Proof. See APPENDIX A.2. □

Remark 5. *We recover Theorem 1 if* $\mathbb{P}_{pred}^A = \mathbb{P}_{pred}$ *in Theorem 3, and we recover Theorem 2 if* $\mathbb{P}_{obs}^A = \mathbb{P}_{obs}$ *in Theorem 4.*

3. STABILITY AND CONVERGENCE USING F-DIVERGENCES

Many common approaches for quantifying the discrepancy between two probability measure are derived from f-divergences. While f-divergences are generally not metrics due to a lack of symmetry, the generalization of the stability results from the total variation metric to f-divergences are relatively straightforward. Below, we provide the formal definition of an f-divergence and provide some context and a brief literature review for the practical application of f-divergences.

Definition 3.1.

Let $\mathbb{P}^A$ and $\mathbb{P}^B$ be probability measures on measure space $(X, \mathcal{B}_X, \mu_X)$ admitting densities π^A and π^B. The

f-divergence is defined as

$$D_f(\mathbb{P}^A \mid\mid \mathbb{P}^B) = \int_{\mathcal{X}} f\left(\frac{\pi^A(x)}{\pi^B(x)}\right)\pi^B(x)d\mu_{\mathcal{X}} \tag{10}$$

where f is a specific convex function defining the f-divergence such that $f(t)$ is bounded $\forall t > 0$, $f(1) = 0$, and $f(0) = \lim_{t^+\to\infty} f(t)$.

In the context of density estimation, f-divergences are often useful in determining optimal parameters or hyper-parameters of a density model [7,27]. For instance, the Kullback-Liebler (KL) divergence can be written as the sum of the negative, expected loglikelihood that the data came from the approximate distribution plus an entropy term independent of the hyper-parameters. Thus, optimal hyper-parameters can be computed by maximizing the loglikelihood, which will then minimize the KL divergence.

Note that in the definition of the f-divergence, we do not necessarily assume that $\mathbb{P}^B$ is absolutely continuous with respect to $\mathbb{P}^A$. If these measures do not possess this property, then the f-divergence is typically defined as infinite, which is not very useful in terms of stability or convergence, so practically we only apply f-divergences to measures that satisfy this absolute continuity condition. When measuring the f-divergence of a measure $\mathbb{P}^B$ from another measure $\mathbb{P}^A$, we write the forward f-divergence as $D_f(\mathbb{P}^A \mid\mid \mathbb{P}^B)$. When the roles of the target and approximate are reversed, i.e., $D_f(\mathbb{P}^B \mid\mid \mathbb{P}^A)$, we call this the *reverse f-divergence*. Note that the reverse f-divergence is not necessarily the same as the forward f-divergence.

The choice of f defines the type of divergence. For instance, choosing $f(t) = \frac{1}{2}|t-1|$ recovers the total variation metric while $f(t) = t\ln t$ defines the KL divergence. KL divergences have found extensive applications in statistics and machine learning, particularly in variational inference [7], optimal experimental design [6,29], and information geometry [2]. Moreover, this particular divergence has served as a useful tool to quantify the information gained in moving from initial to updated measures in data consistent inversion [14,49]. Additionally, it enables the assessment of the distance between the initial and updated densities in terms of the distance between observed and predicted measures, as demonstrated in [12]. Below, we show that this utility can be extended to other types of f-divergences.

3.1 Equivalence of f-divergences within the DCI Framework

Due to the fact that the solution of the stochastic inverse problem is a pullback probability measure, we can make precise statements regarding the f-divergences between measures on the parameter space and corresponding measures on the data space. The following theorem states that the f-divergence of the updated density from the initial density is equal to the f-divergence of the observed from the predicted.

Theorem 6 (f-divergence and DCI)**.** *Given probability measures, $\mathbb{P}_{init}$, $\mathbb{P}_{obs}$, and $\mathbb{P}_{pred}$ which satisfy the Assumption 1 and updated measure $\mathbb{P}_{up}$ given by (4),*

$$D_f\left(\mathbb{P}_{up}\|\mathbb{P}_{init}\right) = D_f\left(\mathbb{P}_{obs}\|\mathbb{P}_{pred}\right).$$

Proof. See APPENDIX B.1. □

In this case, Assumption 1 guarantees that the f-divergence is finite since the observed and updated measures are absolutely continuous with respect to the predicted and initial measures, respectively. The implication is that by computing the f-divergence of relevant measures in the data space $\mathcal{D}$, we obtain the value of the f-divergence of relevant measures in the parameter space Λ and vice-versa. This is valuable when the densities in one space are simpler to evaluate than in another, e.g., if the dimension of one space is smaller or if the densities in one space are given analytically. Next, we consider f-divergences between different updated densities.

3.2 Stability of Updated Densities using f-divergences

The goal of this section is to show stability of densities using an f-divergence in the data space leads to stability of the updated densities on the parameter space. First, we show that the forward f-divergence between two updated densities obtained from the same observed but with different predicted densities is bounded above by a constant times the *reverse* f-divergence between the predicted densities. As the proof demonstrates, the dependence on the reverse f divergence is a consequence of the predicted densities appearing in the denominator of the corresponding updated densities.

Theorem 7 (Predicted Stability in f-divergence)**.** *For fixed measures $\mathbb{P}_{init}$ and $\mathbb{P}_{obs}$ with corresponding densities π_{init} and π_{obs}, respectively, let π^A_{pred} and π^B_{pred} denote predicted densities satsifying Assumption 2.2 and let $\mathbb{P}^A_{up}$ and $\mathbb{P}^B_{up}$ denotes the respective associated updated measures. Additionally, assume there exists another constant $C_1 > 0$ such that*

$$\pi_{pred}(q) \leq C_1 \pi^A_{pred}(q), \quad \text{for a.e. } q \in \mathcal{D}.$$

Then, there exists a constant $C_2 > 0$ such that

$$D_f(\mathbb{P}^A_{up} \,||\, \mathbb{P}^B_{up}) \leq C_2 \cdot D_f(\mathbb{P}^B_{pred} \,||\, \mathbb{P}^A_{pred}).$$

Proof. See APPENDIX B.2. □

Remark 8. *Taking π^A_{pred} to be the push-forward of π_{init}, i.e., $\pi^A_{pred} = \pi_{pred}$, and π^B_{pred} to be some approximation of π_{pred} that converges in the f-divergence, implies convergence of the approximate updated densities in the f-divergence. Note that the additional assumption is trivially satisfied if we take $\pi^A_{pred} = \pi_{pred}$.*

Next, we show that the f-divergence between two updated densities is precisely the f-divergence between the two respective observed densities.

Theorem 9 (Observed Stability in f-divergence)**.** *For fixed measures $\mathbb{P}_{init}$ and $\mathbb{P}_{pred}$ with corresponding densities π_{init} and π_{pred}, respectively, let $\mathbb{P}^A_{obs}$ and $\mathbb{P}^B_{obs}$ denote observed measures satisfying Assumption 2.1 and let $\mathbb{P}^A_{up}$ and $\mathbb{P}^B_{up}$ denote the respective associated updated measures. Then,*

$$D_f(\mathbb{P}^A_{up} \,||\, \mathbb{P}^B_{up}) = D_f(\mathbb{P}^A_{obs} \,||\, \mathbb{P}^B_{obs}).$$

Proof. See APPENDIX B.3. □

Remark 10. *Taking $\pi^A_{obs} = \pi_{obs}$ and π^B_{obs} to be some approximation of π_{obs} that converges in the f-divergence, implies convergence of the approximate updated densities in the f-divergence.*

4. STABILITY AND CONVERGENCE USING INTEGRAL PROBABILITY METRICS (IPM)

Integral probability metrics (IPMs) have become increasingly popular tools in the context of machine learning and generative AI, e.g., see [3,32,34,50]. These metrics are used during the training of neural networks to stabilize the learning process by constraining the generative probability distribution to be similar to the target observed distribution. The class of IPMs includes the maximum-mean-discrepancy [26] and the earth mover's distance [37], among others. We give the abstract definition of an integral probability metric and follow-up with specific cases.

Definition 4.1.

Let $\mathbb{P}^A$ and $\mathbb{P}^B$ be two probability measures on a measure space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$. An integral probability metric is defined as

$$d_{\mathcal{F}}(\mathbb{P}^A, \mathbb{P}^B) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f d\mathbb{P}^A - \int_{\mathcal{X}} f d\mathbb{P}^B \right|$$

where $\mathcal{F}$ is a defined class of real-valued, bounded measureable functions on $\mathcal{X}$.

IPMs generalize certain probability metrics through the appropriate choice of functions $\mathcal{F}$. For instance, by choosing $\mathcal{F}$ to be $\{f : ||f||_\infty \leq 1\}$, where $||f||_\infty$ is the supremum of $|f(x)|$ over $\mathcal{X}$, the resulting metric $d_{\mathcal{F}}$ is equivalent to the total variation metric (see APPENDIX C). The Kantorovich metric, which is the dual of the Wasserstein distance, is obtained by choosing $\mathcal{F} = \{f : ||f||_L \leq 1\}$, where $||f||_L$ is the Lipschitz semi-norm on a metric space $(\mathcal{X}, \rho)$,

$$||f||_L := \sup \left\{ \frac{|f(x) - f(y)|}{\rho(x, y)} \ : \ x \neq y \text{ in } \mathcal{X} \right\}.$$

The Kernel distance or maximum mean discrepancy is obtained when $\mathcal{F} = \{f : ||f||_{\mathcal{H}} \leq 1\}$, where $\mathcal{H}$ represents a reproducing kernel Hilbert space.

4.1 Using IPM within the DCI Framework

In the context of DCI, we are quantifying distances between measures on different spaces Λ and $\mathcal{D}$. It is therefore appropriate to consider IPMs defined by different function spaces. Specifically, we consider a family of functions $\mathcal{F}_\Lambda$ be a set of real valued functions $\{f : \Lambda \to \mathbb{R}\}$ and a set $\mathcal{G}_{\mathcal{D}}$ where $\{g : \mathcal{D} \to \mathbb{R}\}$. These two families may, in general, reproduce the same norms (as is the case when $\mathcal{F}_\Lambda$ and $\mathcal{G}_{\mathcal{D}}$ are chosen to induce total variation metrics), but this is not necessary.

Our goal is to establish the relationship between the metrics defined by two function space $\mathcal{F}_\Lambda$ and $\mathcal{G}_{\mathcal{D}}$, examining how approximations of measures in the data space impact the corresponding updated measures in the parameter space. As in the coming analysis of stability in L^p metrics in Section 5, the ratio of π_{init} and π_{pred} plays a critical role.

Now consider the IPM defined by $\mathcal{F}_\Lambda$ between two updated densities with different observed densities π^A_{obs} and π^B_{obs},

$$d_{\mathcal{F}_\Lambda}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) = \sup_{f \in \mathcal{F}_\Lambda} \left| \int_\Lambda f(\lambda)(\pi^A_{\text{up}}(\lambda) - \pi^B_{\text{up}}(\lambda)) \, d\mu_\Lambda \right|.$$

Applying the disintegration theorem and Definition 2.3 gives

$$d_{\mathcal{F}_\Lambda}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) = \sup_{f \in \mathcal{F}_\Lambda} \left| \int_{\mathcal{D}} \left(\int_{\Lambda \cap Q^{-1}(q)} f(\lambda) \pi_{\text{init}\,|q}(\lambda)\, d\mu_{\Lambda,q} \right) (\pi^A_{\text{obs}}(q) - \pi^B_{\text{obs}}(q))\, d\mu_{\mathcal{D}} \right|$$

The inner integral is the expected value of the function $f \in \mathcal{F}_\Lambda$ conditioned on q, i.e.,

$$\mathbb{E}_{\Lambda|q}(f(\lambda)) = \int_{\Lambda \cap Q^{-1}(q)} f(\lambda) \pi_{\text{init}\,|q}(\lambda) d\mu_{\Lambda,q}$$

where $\mathbb{E}_{\Lambda|q}$ is the expected value taken with respect to the conditional initial measure $\mathbb{P}_{\text{init}\,|q}$. We note that $\mathbb{E}_{\Lambda|q}$ is a linear operator acting on the space of functions $\mathcal{F}_\Lambda$ and mapping them to the space of functions $\mathcal{G}_{\mathcal{D}}$. Thus, a sufficient condition for determining the stability of the updated density with respect to the observed or predicted distribution using different integral probability metrics is that $\mathbb{E}_{\Lambda|q}$ is a bounded operator. The following theorem shows how to relate the two metrics defined by $\mathcal{F}_\Lambda$ and $\mathcal{G}_{\mathcal{D}}$ to ensure that convergence of an approximate observed or predicted distribution in the data space will guarantee convergence of the approximate updated distribution in the parameter space.

Theorem 11 (Predicted Stability in IPM)**.** *Let $\mathcal{F}_\Lambda$ and $\mathcal{G}_{\mathcal{D}}$ be used to define IPM for measures on Λ and $\mathcal{D}$, respectively. Suppose $\mathbb{E}_{\Lambda|q}$ is a bounded operator from $\mathcal{F}_\Lambda$ to $\mathcal{G}_{\mathcal{D}}$. For fixed measures $\mathbb{P}_{init}$ and $\mathbb{P}_{obs}$ with corresponding densities π_{init} and π_{obs} respectively, let π^A_{pred} and π^B_{pred} denote predicted densities satisfying Assumption 2.2 and let $\mathbb{P}^A_{up}$ and $\mathbb{P}^B_{up}$ denotes the respective associated updated measures. Additionally, assume there exists another constant $C_1 > 0$ such that*

$$\pi_{pred}(q) \leq C_1 \pi^A_{pred}(q), \quad \text{for a.e. } q \in \mathcal{D}.$$

Then, there exists a constant $C_2 > 0$ such that

$$d_{\mathcal{F}_\Lambda}(\mathbb{P}^A_{up}, \mathbb{P}^B_{up}) \leq C_2 d_{\mathcal{G}_{\mathcal{D}}}(\mathbb{P}^A_{pred}, \mathbb{P}^B_{pred})$$

Proof. See APPENDIX D.1. □

Theorem 12 (Observed Stability in IPM)**.** *Let $\mathcal{F}_\Lambda$ and $\mathcal{G}_{\mathcal{D}}$ be used to define IPM for measures on Λ and $\mathcal{D}$, respectively. Suppose $\mathbb{E}_{\Lambda|q}$ is a bounded operator from $\mathcal{F}_\Lambda$ to $\mathcal{G}_{\mathcal{D}}$. For fixed measures $\mathbb{P}_{init}$ and $\mathbb{P}_{pred}$ with corresponding densities π_{init} and π_{pred} respectively, let $\mathbb{P}^A_{obs}$ and $\mathbb{P}^B_{obs}$ denote observed measures satisfying Assumption 2.1 and let*

$\mathbb{P}_{up}^A$ and $\mathbb{P}_{up}^B$ denote the respective associated updated measures. Then, there exists $C > 0$ such that

$$d_{\mathcal{F}_\Lambda}(\mathbb{P}_{up}^A, \mathbb{P}_{up}^B) \leq C d_{\mathcal{G}_\mathcal{D}}(\mathbb{P}_{obs}^A, \mathbb{P}_{obs}^B)$$

Proof. See APPENDIX D.2. □

Remark 13. *Similar to Remarks 10 and 8, we can take $\pi_{obs}^A = \pi_{obs}$ (or $\pi_{pred}^A = \pi_{pred}$), and if we assume π_{obs}^B converges to π_{obs} (or π_{pred}^B converges to π_{pred}) in the IPM, then the approximate updated densities also converge in the appropriate IPM.*

Theorems 11 and 12 provide sufficient conditions for determining stability using IPMs that involve the boundedness of $\mathbb{E}_{\Lambda|q}$. In some cases, it is straightforward to verify this condition holds. For instance, if $\mathcal{F}_\Lambda$ is defined as $\{f : (||f||_\infty + ||f||_L) \leq 1\}$, then $\mathcal{F}_\Lambda$ induces the so-called Dudley metric. If we compare this to $\mathcal{G}_\mathcal{D}$ defined as the total variation metric, i.e., $\{g : ||g||_\infty \leq 1\}$ and Λ is compact (and finite dimensional), we can show that $\mathbb{E}_{\Lambda|q}$ is bounded since, $\forall f$ and $\forall q$,

$$\begin{aligned} \left|\mathbb{E}_{\Lambda|q}(f)\right| &= \left|\int_{\Lambda\cap Q^{-1}(q)} f(\lambda)\pi_{\text{init}\,|q}(\lambda)d\mu_{\Lambda,q}\right| \\ &\leq \left| ||f||_\infty \cdot \int_{\Lambda\cap Q^{-1}(q)} \pi_{\text{init}\,|q}(\lambda)d\mu_{\Lambda,q}\right| \\ &= ||f||_\infty \leq ||f||_\infty + ||f||_L, \end{aligned}$$

which implies that

$$||\mathbb{E}_{\Lambda|q}(f)||_{\mathcal{G}_\mathcal{D}} = ||\mathbb{E}_{\Lambda|q}(f)||_\infty \leq ||f||_\infty + ||f||_L = ||f||_{\mathcal{F}_\Lambda}.$$

It is worth noting that while this condition is sufficient for determining stability using integral probability metrics, it is not necessary. Indeed, in [54], it is shown that as long as there exists functions $g \in \mathcal{G}_\mathcal{D}$ that can dominate functions $\mathbb{E}_{\Lambda|q} f$ in a piecewise-sense, then the stability condition holds.

4.2 A pullback IPM

We close this section with a concise description of how to construct an IPM on the parameter space that is equal to a given IPM on the data space. This has potential applications in settings where machine learning algorithms are used to produce observed distributions that rely on optimizing an unorthodox IPM $\mathcal{F}_\mathcal{D}$.

This is also practical when the goal is to generate approximated updated distributions that are close to an exact updated distribution based precisely on how close the associated approximate observed distribution is to an exact observed distribution.

Definition 4.2.

Let $d_{\mathcal{F}_\mathcal{D}}$ be an IPM on the distributions of the data space defined by $\mathcal{F}_\mathcal{D}$. Let Q be a measurable quantity of interest map $Q : \Lambda \to \mathcal{D}$. Define a class of functions $\mathcal{F}^*_\Lambda$ such that for every $g \in \mathcal{F}_\mathcal{D}$

$$f(\lambda) := (g \circ Q)(\lambda) = g(Q(\lambda)).$$

Then, we define the pullback IPM with respect to $d_{\mathcal{F}_\mathcal{D}}$ as

$$d_{\mathcal{F}^*_\Lambda}(\mathbb{P}^A, \mathbb{P}^B) = \sup_{f \in \mathcal{F}^*_\Lambda} \left| \int_\Lambda f d\mathbb{P}^A - \int_\Lambda f d\mathbb{P}^B \right|$$

We can verify that this definition produces a valid integral probability metric by recalling the assumption that Q and g are measurable functions on corresponding Borel sets. Since $Q : (\Lambda, \mathcal{B}_\Lambda) \to (\mathcal{D}, \mathcal{B}_\mathcal{D})$ is measurable and $g : (\mathcal{D}, \mathcal{B}_\mathcal{D}) \to (\mathbb{R}, \mathcal{B})$ is measurable for all $g \in \mathcal{F}_\mathcal{D}$, the composition $f = g \circ Q$ is measurable. Also, since every $g \in \mathcal{F}_\mathcal{D}$ is bounded so too must each $f \in \mathcal{F}^*_\Lambda$, thus satisfying the definition of an IPM.

The next theorem shows that the pullback IPM measures differences between updated densities by the differences between their corresponding distributions in the data space, i.e. how different are the push-forwards with respect to an IPM on $\mathcal{D}$.

Theorem 14 (Stability using the Pullback IPM)**.** *For fixed measures $\mathbb{P}_{init}$ and $\mathbb{P}_{obs}$ with corresponding densities π_{init} and π_{obs} respectively, let π^A_{pred} and π^B_{pred} denote predicted densities satisfying Assumption 2.2 and let $\mathbb{P}^A_{up}$ and $\mathbb{P}^B_{up}$ denotes the respective associated updated measures. Additionally, assume there exists another constant $C_1 > 0$ such that*

$$\pi_{pred}(q) \le C_1 \pi^A_{pred}(q), \quad \text{for a.e. } q \in \mathcal{D}.$$

Then, there exists a constant $C_2 > 0$ such that

$$d_{\mathcal{F}^*_\Lambda}(\mathbb{P}^A_{up}, \mathbb{P}^B_{up}) \le C_2 d_{\mathcal{F}_\mathcal{D}}(\mathbb{P}^A_{pred}, \mathbb{P}^B_{pred}). \tag{11}$$

Similarly, for fixed measures $\mathbb{P}_{init}$ and $\mathbb{P}_{pred}$ with corresponding densities π_{init} and π_{pred} respectively, let $\mathbb{P}^A_{obs}$ and $\mathbb{P}^B_{obs}$

denote observed measures satisfying Assumption 2.1 and let $\mathbb{P}_{up}^A$ and $\mathbb{P}_{up}^B$ denote the respective associated updated measures. Given an IPM on $\mathcal{D}$ defined by $\mathcal{F}_\mathcal{D}$ and the corresponding data-consistent IPM defined by $\mathcal{F}_\Lambda^*$, we have

$$d_{\mathcal{F}_\Lambda^*}(\mathbb{P}_{up}^A, \mathbb{P}_{up}^B) = d_{\mathcal{F}_\mathcal{D}}(\mathbb{P}_{obs}^A, \mathbb{P}_{obs}^B), \tag{12}$$

Proof. See APPENDIX D.3. □

5. CONVERGENCE OF UPDATED DENSITIES IN L^P

This section focuses on convergence of the updated density in L^p-metrics. Due to the complexity of some of the technical details in this section, we do not pursue the more general scenario considered in previous sections with (somewhat) arbitrary π_{up}^A and π_{up}^B, and focus on the special case where $\pi_{\text{up}}^A = \pi_{\text{up}}$ and π_{up}^B involves an approximation.

While TV is a commonly used metric for evaluating density estimations, the mean-integrated squared error or MISE is perhaps the dominant metric considered within the kernel density estimation literature. This is equivalent to measuring the mean L^2-error (squared) between distributions, and is therefore also referred to as the L^2-risk. Other density estimation techniques use more general L^p-risk to prove various theoretical convergence results and to determine bounds on the rate of convergence [21,51]. Given these considerations, we seek to generalize Theorems 1 and 2 to the general class of L^p metrics with $p > 1$, which, as we illustrate, are more difficult to work with than the total variation metric. To be precise, we aim to show that the convergence of any sequence of approximations $\pi_{\text{pred}}^n \to \pi_{\text{pred}}$ or $\pi_{\text{obs}}^n \to \pi_{\text{obs}}$ that converges in L^p implies the convergence of the updated densities $\pi_{\text{up}}^n \to \pi_{\text{up}}$ in L^p. First, we define the L^p metric measuring the difference between two probability measures.

Definition 5.1.

Let $\mathbb{P}^A$ and $\mathbb{P}^B$ be two probability measures on a measure space $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mu_\mathcal{X})$ admitting densities π^A and π^B. Then the L^p-metric (or distance) over $\mathcal{X}$ between $\mathbb{P}^A$ and $\mathbb{P}^B$ is defined as

$$d_{L^p(\mathcal{X})}(\mathbb{P}^A, \mathbb{P}^B) := \left(\int_\mathcal{X} \left| \pi^A(x) - \pi^B(x) \right|^p d\mu_\mathcal{X} \right)^{1/p} = \left\| \pi^A - \pi^B \right\|_{L^p(\mathcal{X})}$$

for any $1 \le p < \infty$.

Note that if $p = 1$, then this reduces to the TV metric given in Definition 2.4.

5.1 Rates of Convergence

We are also interested in analyzing the *rate of convergence* of the approximation to the updated density in relation to rates of convergence of the density approximations to either π_{pred} or π_{obs} in $\mathcal{D}$. The results we obtain in this section show that the rate of convergence is of the same order on *almost* all of Λ but not necessarily on *all* of Λ. This is similar in spirit to other convergence proofs of density estimates which are shown to hold true on sequences of nested compact sets that converge from below to the full domain, e.g., see [28]. It is also common for results in measure theory to refer to a property holding everywhere except on a measurable set of arbitrarily small size, e.g., see Luzin's theorem (cf. Theorem 7.10 in [23]) and Egoroff's theorem (cf. Theorem 2.33 in [23]). We define rate of convergence in an almost sense formally below.

Definition 5.2.

Let $\mathbb{P}^n$ be a sequence of probability measures on measure space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mu_{\mathcal{X}})$ which converges to $\mathbb{P}$ in metric $d_{L^p(\mathcal{X})}$ defined over the domain $\mathcal{X}$. We say the convergence rate of $\mathbb{P}^n$ is of order $O(\rho(n))$ in an almost sense if for every $\epsilon > 0$ there exists a measurable subset A of $\mathcal{X}$ such that $\mathbb{P}(A) < \epsilon$ and

$$d_{L^p(\mathcal{X}\setminus A)}(\mathbb{P}^n, \mathbb{P}) \leq M\rho(n) \quad \forall n \geq N \tag{13}$$

for some $M, N \in \mathbb{R}$.

Practically, Definition 5.2 implies that this order of convergence holds on "most" of the space since $\mathbb{P}(\mathcal{X} \setminus A) \geq 1 - \epsilon$. For example, if $\epsilon = 0.01$, we can guarantee the existence of a set that is at least 99% probable such that that the order of convergence holds on this set. Note that because $\mathbb{P}^n \to \mathbb{P}$ in $L^p(\mathcal{X})$, $\mathbb{P}^n$ still converges to $\mathbb{P}$ on the "small set" A, the definition simply states that the convergence rate is something other than $O(\rho(n))$ on this small set. Indeed, since ϵ is arbitrary, we can make $\mathcal{X}\setminus A$ as close to $\mathcal{X}$ in measure $\mathbb{P}$ as is desired, hence the use of the term "almost all" of $\mathcal{X}$. With the above two definitions, we proceed to analyzing the convergence and rate of convergence of the updated density in terms of the L^p-metric over Λ.

Note that both Theorem 1 and 2 require their own versions of the predictability assumption, which is necessary to guarantee existence of the solution to the inverse problem using either approximation. In this paper, we are primarily interested in a more general case where it is possible to define a sequence of approximations that converge in L^p. As in [16], using approximations of the observed or predicted densities requires the assumption that, in an asymptotic sense, these approximations satisfy versions of the predictability assumption to guarantee the existence of solutions to the inverse problem using these approximations. For convenience, we combine these two cases in the following assumption that includes a third case involving simultaneous approximations of both the observed and predicted densities, which is a common occurrence in practice.

Assumption 3.

There exists a constant $C > 0$ such that:

1. Given a sequence of approximate observed densities, (π^m_{obs}), there exists an M such that $\forall m \geq M$,

$$\pi^m_{\text{obs}}(q) \leq C\pi_{\text{pred}}(q) \quad a.e.\ q \in \mathcal{D}. \tag{14}$$

2. Given a sequence of approximate predicted densities, $\left(\pi^n_{\text{pred}}\right)$, there exists an N such that $\forall n \geq N$,

$$\pi_{\text{obs}}(q) \leq C\pi^n_{\text{pred}}(q) \quad a.e.\ q \in \mathcal{D}. \tag{15}$$

3. Given sequences of approximate observed densities and predicted densities, which satisfy (14) and (15) there exists a K such that $\forall m, n \geq K$ we have

$$\pi^m_{\text{obs}}(q) \leq C\pi^n_{\text{pred}}(q) \quad a.e.\ q \in \mathcal{D}. \tag{16}$$

The following corollaries describe rates of convergence in $L^1(\Lambda)$ of updated density approximations.

Corollary 1. *If $\pi^m_{obs} \to \pi_{obs}$ in $L^1(\mathcal{D})$ with rate of convergence $O(\rho(m))$ and Assumption 3.1 is satisfied, then $\pi^m_{up} \to \pi_{up}$ in $L^1(\Lambda)$ with rate of convergence $O(\rho(m))$.*

Proof. The proof is an immediate consequence of Theorem 2. □

Corollary 2. *If $\pi^n_{pred} \to \pi_{pred}$ in L^1 with rate of convergence $O(\rho(n))$ and Assumption 3.2 is satisfied, then $\pi^n_{up} \to \pi_{up}$ in $L^1(\Lambda)$ with rate of convergence $O(\rho(n))$.*

Proof. The proof is an immediate consequence of Theorem 1. □

Corollary 3. *If $\pi^n_{pred} \to \pi_{pred}$ and $\pi^m_{obs} \to \pi_{obs}$ in $L^1(\mathcal{D})$ with rates of convergence $O(\rho(n))$ and $O(\gamma(m))$, respectfully, and Assumption 3.3 is satisfied, then $\pi^n_{up}\pi_{up}$ in $L^1(\Lambda)$ with rate of convergence $O(\rho(n) + \gamma(m))$.*

Proof. The proof follows from applying a triangle inequality to the TV metric. □

5.2 Stability and Convergence in L^p with Approximate Densities

First, we show that the updated density converges to the true updated density in the L^p-metric on Λ if the approximation of the predicted density converges in the L^p-metric on $\mathcal{D}$. It is worth noting that an additional assumption involving the initial density belonging to L^∞ is made to avoid singularities that complicate the proofs. Since we are typically free to choose initial densities in the setup of the problems, this is often a trivial assumption to satisfy in practice.

Theorem 15 (L^p Convergence with Approximated Predicted Densities). *Suppose $\pi_{init} \in L^\infty(\Lambda)$ and π_{obs} are chosen so that Assumption 1 is satisfied. If (π^n_{pred}) satisfies Assumption 3.2 and $\pi^n_{pred} \to \pi_{pred}$ in $L^p(\mathcal{D})$, then $\pi^n_{up} \to \pi_{up}$ in $L^p(\Lambda)$.*

Proof. See APPENDIX E.1. □

Theorem 16 (Rate of Convergence with Predicted in L^p). *Suppose $\pi_{init} \in L^\infty(\Lambda)$ and π_{obs} are chosen so that Assumption 1 is satisfied. If (π^n_{pred}) satisfies Assumption 3.2, $\pi^n_{pred} \to \pi_{pred}$ in $L^p(\mathcal{D})$, and the convergence rate of $\mathbb{P}^n_{pred}$ is of order $O(\rho(n))$ on almost all of $\mathcal{D}$, then the convergence rate of $\mathbb{P}^n_{up}$ is of order $O(\rho(n))$ on almost all of Λ.*

Proof. See APPENDIX E.2. □

Next, we show that the updated density converges to the true updated density in the L^p-metric on Λ if the approximation of the observed density converges in the L^p-metric on $\mathcal{D}$.

Theorem 17 (L^p Convergence with Approximated Observed Densities). *Suppose $\pi_{init} \in L^\infty(\Lambda)$ and π_{obs} are chosen so that Assumption 1 is satisfied. If (π^n_{obs}) satisfies Assumption 3.1 and $\pi^n_{obs} \to \pi_{obs}$ in $L^p(\mathcal{D})$, then $\pi^n_{up} \to \pi_{up}$ in $L^p(\Lambda)$.*

Proof. See APPENDIX E.3. □

Theorem 18 (Rate of Convergence with Observed in L^p). *Suppose $\pi_{init} \in L^\infty(\Lambda)$ and π_{obs} are chosen so that Assumption 1 is satisfied. If (π^n_{obs}) satisfies Assumption 3.1, $\pi^n_{obs} \to \pi_{obs}$ in $L^p(\mathcal{D})$, and the convergence rate of $\mathbb{P}^n_{obs}$ is of order $O(\rho(n))$ on almost all of $\mathcal{D}$, then the convergence rate of $\mathbb{P}^n_{up}$ is of order $O(\rho(n))$ on almost all of Λ.*

Proof. See APPENDIX E.4. □

Theorems 16 and 18 demonstrate that in the L^p-metric, the order of convergence for the updated density is equal to the order of convergence for the density approximations in the data space. While these theorems are not quite as strong as their counterparts in the total variation metric, they are more generally applicable to common measures of convergence, especially the mean-integrated squared error (MISE) or L^2-risk. In addition, similar to the L^1 results, these theorems imply that, as long as the dimension of the data space is relatively small, the curse of dimensionality associated with estimating the updated density in the parameter space can be mitigated, which is beneficial if the dimension of the parameter space is large since many density estimation techniques scale poorly with dimension. We conclude this section by noting that a simple application of the triangle inequality can be used for the case where *both* the observed and predicted densities are approximated if Assumption 3.3 is satisfied.

6. NUMERICAL EXAMPLES

The examples of this section are intended to be straightforward and reproducible to highlight key aspects of the theoretical results presented above in the context of practical approximation issues.

6.1 Estimating Discrepancies in Parameter Space via Discrepancies in Data Space

This example illustrates how f-divergences are useful in the context of practical approximation issues encountered when constructing a kernel density estimate (KDE) in the DCI framework. We focus on numerically demonstrating Theorem 9, which gives

$$D_f(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) = D_f(\mathbb{P}^A_{\text{obs}}, \mathbb{P}^B_{\text{obs}}).$$

In the interest of space, we limit the presentation to the KL divergence, which is a commonly used f-divergence. We demonstrate how the above result is useful in quantifying the discrepancy between distinct updated densities in the parameter space by measuring the discrepancy between the associated estimates of observed densities in the data space obtained via different bandwidth parameter selection techniques utilized in the KDE estimates. For the interested reader, the supplemental files (see APPENDIX E.5) provides the code to generate both the results presented here as well as additional results that utilize Gaussian Mixture Models (GMMs), which are popular semi-parametric density estimation techniques. These additional results include numerical demonstration of the thematically similar theoretical results from both

Section 3 and Section 4 that relate discrepancies (measured in either f-divergences or IPMs) between data space densities (whether observed or predicted densities) to discrepancies in the associated updated densities in the parameter space.

6.1.1 DCI Setup

Consider the QoI map $Q : \Lambda \to \mathcal{D}$ given by

$$Q(\lambda) = \begin{bmatrix} \lambda_1 \cos \lambda_2 \\ \lambda_1 \sin \lambda_2 \end{bmatrix} + \begin{bmatrix} \lambda_3 \\ \lambda_4 \end{bmatrix}.$$

This mapping draws circular arcs of radius λ_1 and angle λ_2 around a central point (λ_3, λ_4). Suppose that data points are randomly drawn from circular arcs centered around points three uncertain points $\{\mu_k\}_{k=1}^3 \subset \mathbb{R}^2$. We consider these central points to be uncertain as well as the sampling distribution of their radii and arc lengths, and represent the initial state of uncertainty as:

$$\lambda_1 \sim U[0.65, 1.35],\ \lambda_2 \sim U[0, 2\pi],\ (\lambda_3, \lambda_4) \sim \sum_{k=1}^{3} w_k \mathcal{N}(\mu_k, \sigma^2 I),$$

where $w_k = \frac{1}{3}$ are the weights of a mixture normal distribution with means located at the centers $\mu_k = \{(-1, 0.5), (0, 0), (1, 0.5)\}$ and variance determined by $\sigma^2 = 0.005$. The left plot in Figure 1 shows the corresponding push-forward sample of $m = 15000$ predicted points $q \in \mathcal{D}$ drawn from these initial distributions.

The right plot in Figure 1 shows $n = 500$ observations drawn from the so-called "dual moons" dataset, which is a commonly utilized dataset used for evaluating density estimation in machine learning. The DCI problem in this example is to utilize Q to find an updated probability density on λ that is consistent with estimated densities computed from this observed "dual moons" dataset.

6.1.2 KDEs and the Bandwidth Parameter

Given a sample $x_1, \ldots, x_n$ from an unknown distribution π, the KDE is defined by,

$$\pi_{\text{KDE}}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

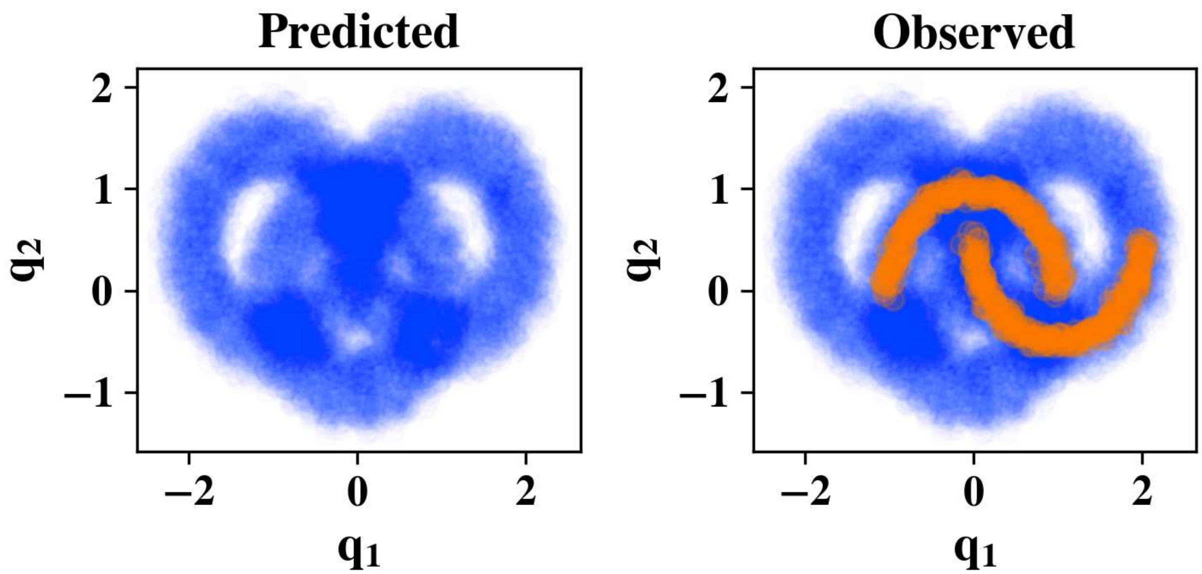

FIG. 1: Predicted QoI (left). Observed QoI shown superimposed on the predicted QoI (right).

where K is a non-negative kernel function and h is a bandwidth parameter. Perhaps the most commonly used kernel is a Gaussian kernel, and the resulting Gaussian KDE is simply referred to as GKDE. We use the GKDE in this example.

One of the challenges with using a KDE (regardless of the kernel choice) is determining an appropriate bandwidth parameter h for the density approximation. If the bandwidth is chosen too large, the resulting density is over-smoothed, but if the bandwidth is chosen too small, the density overfits the data resulting in increased variance in the estimate between different sample sets of the same size.

Heuristic approaches for choosing the bandwidth are common, but they often make strong assumptions about the target density. For instance, Silverman's rule-of-thumb [41] gives

$$h_{silver} := \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{\frac{-1}{d+4}} \hat{\sigma}_q, \tag{17}$$

where $\hat{\sigma}_q$ is the computed variance of the sample data. This heuristic is based on an assumption that the samples are independently and identically distributed from a normal distribution.

Alternatively, a statistical strategy, such as cross-validation [52], can optimally choose the bandwidth with respect to some criteria, e.g., the KL divergence. Minimizing the KL divergence is equivalent to maximizing the expected loglikelihood of the data, which makes it easy to compute, i.e.,

$$h_{cv} := \arg\min_{h} D_{KL}(\mathbb{P}_{\text{obs}}, \widetilde{\mathbb{P}}_{\text{obs}}) = \arg\max_{h} \mathbb{E}_{q\sim\pi_{\text{obs}}} \left[\log \pi_{\text{KDE}}(q)\right]. \tag{18}$$

6.1.3 Quantifying Impact of Bandwidth Selection

Fig. 2 compares results using either (17) or (18) to construct a GKDE estimate of the observed density. The use of h_{silver} clearly oversmooths the GKDE compared to the use of h_{cv}. The impact of these distinct estimates of the observed density on the corresponding updated densities is shown in Fig. 3, where it is clear that the oversmoothed observed density leads to an insignificant update to the marginal of λ_1.

We can quantitatively measure the impact of the choice of bandwidths on the updated density—or in this example, the information gained from choosing h_{cv} instead of h_{silver}—using the KL divergence. Moreover, Theorem 9 states that to measure these differences between updated densities, it suffices to measure the differences between the observed densities in the data space. Using the notation of Theorem 9, denote the GKDEs obtained using h_{silver} and h_{cv} by π^A_{obs} and π^B_{obs}, and the corresponding updated densities by π^A_{up} and π^B_{up}, respectively. We use Monte-Carlo sampling ($M = 10000$ samples) from π^A_{up} and π^A_{obs} to estimate the KL divergences in both the parameter and data spaces. Over $B = 30$ batches, the resulting average estimate of $D_{KL}(\pi^A_{\text{obs}}, \pi^B_{\text{obs}}) \approx 0.953$ with a standard deviation of 0.008. The average estimate of $D_{KL}(\pi^A_{\text{up}}, \pi^B_{\text{up}}) \approx 0.950$ with a standard deviation of 0.006. As expected, the estimates of the KL divergence are nearly equal up to errors due to sampling. This illustrates the utility of Theorem 9: the computation of discrepancies between approximate densities in the parameter space can be replaced by a potentially more efficient computation of discrepancies between approximate densities in the data space where the actual approximations take place. This implies that the discrepancies between updated densities can be estimated without the need to solve the stochastic inverse problem.

6.2 L^p Order of Convergence in an Almost Sense

Here, we numerically demonstrate Theorem 16 that states that the rate of convergence in L^p of approximated updated densities is the same (in the measure-theoretic almost sense) as that of the associated approximated predicted densities.

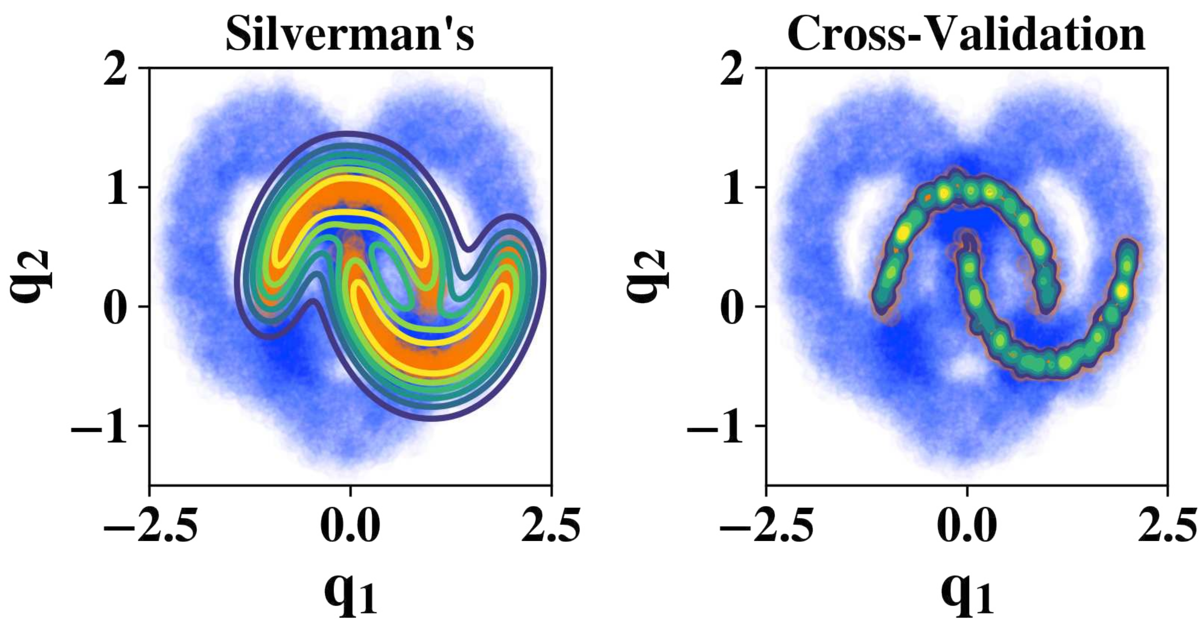

FIG. 2: The left plot shows the approximation of the observed density using a GKDE and Silverman's rule-of-thumb, h_{silver}, for the bandwidth parameter. This clearly leads to an oversmoothed estimate of the density. The right plot shows the approximation of the observed density using a GKDE and cross-validation to select the bandwidth parameter, h_{cv}, which leads to a better estimate of the distribution of the dual moons dataset.

6.2.1 DCI Setup

Consider the linear QoI map $Q : \Lambda \to \mathcal{D}$ from $\mathbb{R}^2$ to $\mathbb{R}$ defined by $Q(\lambda) = \lambda_1 + \lambda_2$, with a triangular observed density defined by:

$$\pi_{\text{obs}}(q) = \begin{cases} 4q & 0 \le q < \frac{1}{2}, \\ -4(q-1) & \frac{1}{2} \le q \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Let the initial density be uniform on $[0, 1] \times [0, 1]$, then the exact predicted density is also a triangular

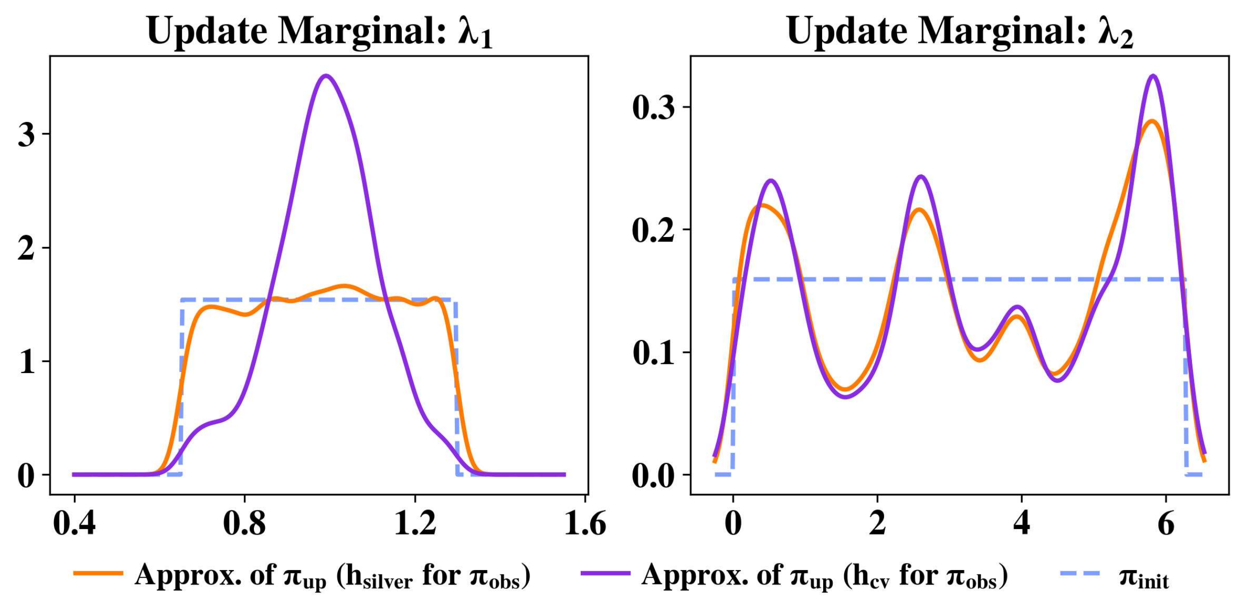

FIG. 3: The initial and approximate updated marginals associated with the first two components of λ: the radius parameter λ_1 and angle parameter λ_2. The right plot shows that the two approximations of the updated density π_{up} (using either h_{silver} or h_{cv} for the GKDE of π_{obs}) have marginals that appear to mostly agree on λ_2 and differ from the marginal of the initial density π_{init}. The left plot shows that the approximation of π_{up} associated with h_{silver} fails to produce an update for λ_1 that is significantly different from π_{init}. On the other hand, the approximation of π_{up} associated with h_{cv} shows a distribution of λ_1 that is not uniformly distributed.

density:

$$\pi_{\text{pred}}(q) = \begin{cases} q & 0 \le q < 1, \\ -(q-2) & 1 \le q \le 2, \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

For the sake of illustration, suppose we approximate this density using the following sequence,

$$\pi^n_{\text{pred}}(q) = \begin{cases} q + \frac{1}{n} & 0 \le q < 1, \\ -(q-2) - g_n(q) & 1 \le q \le 2, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

where $g_n(q) = -\frac{2}{n}(q-2)$ is chosen so that π^n_{pred} is a valid probability distribution that integrates to 1 (true for any $n \ge 2$). Fig. 4 shows the observed, predicted, and approximations of the predicted for $n = 2, 4,$ and 8.

6.2.2 Rates of Convergence

The L^p-error in the approximation of π^n_{pred} has the following closed form,

$$\begin{aligned} ||\pi^n_{\text{pred}} - \pi_{\text{pred}}||_{L^p(\mathcal{D})} &= \left(\int_{\mathcal{D}} \left| \pi^n_{\text{pred}}(q) - \pi_{\text{pred}}(q) \right|^p d\mu_{\mathcal{D}} \right)^{1/p} \\ &= \left(\int_0^1 \left| \frac{1}{n} \right|^p d\mu_{\mathcal{D}} + \int_1^2 \left| \frac{2}{n} \cdot (q-2) \right|^p d\mu_{\mathcal{D}} \right)^{1/p} \\ &= \left(\frac{1}{n^p} + \left(\frac{2^p}{n^p} \right) \left(\frac{1}{p+1} \right) \right)^{1/p} \\ &= \frac{1}{n} \cdot \left(1 + \frac{2^p}{p+1} \right)^{1/p}. \end{aligned}$$

For a fixed p, this clearly converges to 0 with order $O(n^{-1})$.

Since the predictability assumption is satisfied for each $n \ge 2$ (i.e., π^n_{pred} is absolutely continuous with

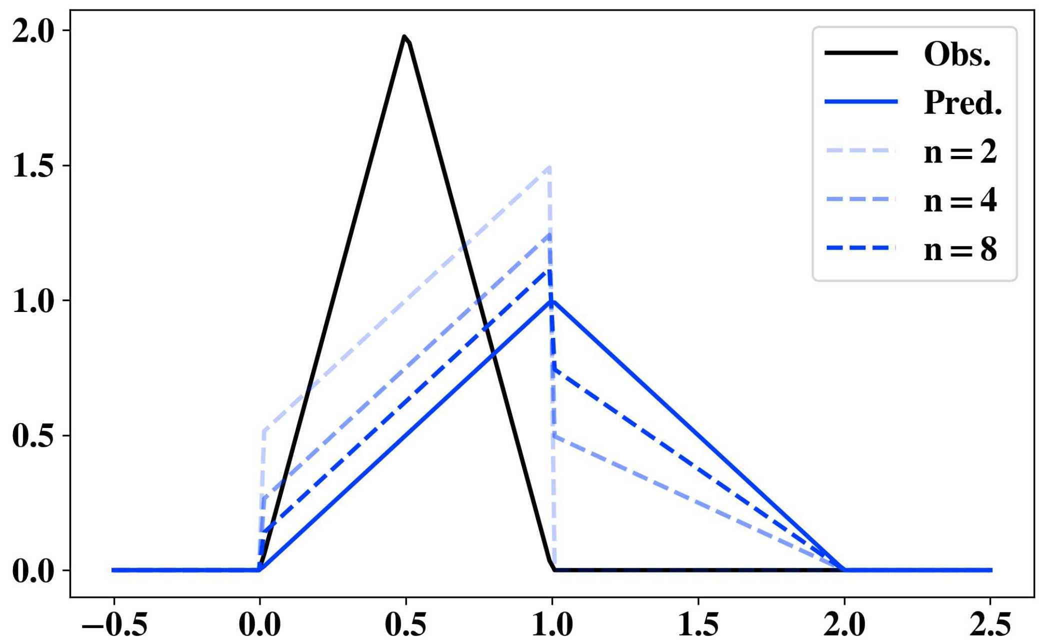

FIG. 4: The observed and predicted densities (solid lines) defined by Eqs. (19) and (20). Dashed lines show a sequence of approximations defined by Eq. (21) that converge to the predicted density.

respect to π_{obs}), we can compute the approximate updated density, π_{up}^n, for each n, as

$$\pi_{\text{up}}^n(\lambda) = \begin{cases} \dfrac{4Q(\lambda)}{Q(\lambda)+\frac{1}{n}} & 0 \le Q(\lambda) < \frac{1}{2} \\ \dfrac{-4(Q(\lambda)-1)}{Q(\lambda)+\frac{1}{n}} & \frac{1}{2} \le Q(\lambda) \le 1 \\ 0 & \text{otherwise} \end{cases}$$

for any $\lambda \in [0,1] \times [0,1]$. As $n \to \infty$, the approximate updated densities converge to the exact updated density, π_{up}, given by

$$\pi_{\text{up}}(\lambda) = \begin{cases} 4 & 0 < Q(\lambda) < \frac{1}{2} \\ \dfrac{-4(Q(\lambda)-1)}{Q(\lambda)} & \frac{1}{2} \le Q(\lambda) \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Theorem 15 guarantees that this convergence is in L^p since $\pi_{\text{pred}}^n \to \pi_{\text{pred}}$ in L^p.

Fig. 5 shows the corresponding convergence rates of $\pi_{\text{pred}}^n \to \pi_{\text{pred}}$ versus $\pi_{\text{up}}^n \to \pi_{\text{up}}$ in L^p with $p = 4$. Note that while $\pi_{\text{up}}^n \to \pi_{\text{up}}$, the order of convergence is closer to $O(n^{-0.65})$ rather than the rate of $\pi_{\text{pred}}^n \to \pi_{\text{pred}}$, which is $O(n^{-1})$. Indeed, for this specific example, we can show that the rate of convergence of the updates must be strictly less than the rate of convergence between the predicted densities in L^p.

However, according to Theorem 16, if we fix an $\epsilon > 0$, there exists a set A_δ such that the $\mathbb{P}_{\text{up}}(A_\delta) < \epsilon$ and the rate of convergence of $\pi_{\text{up}}^n \to \pi_{\text{up}}$ in L^p over $\Lambda \setminus Q^{-1}(A_\delta)$ is $O(n^{-1})$. For this example, if we choose $\delta < \sqrt{\frac{\epsilon}{2}}$ and take the small set A_δ as in the proof of Theorem 18 (see APPENDIX E.1), i.e.,

$$A_\delta := \{q : \pi_{\text{pred}}(q) \le \delta\} = \{q \le \delta\} \cup \{-(\delta - 2) \le q\}, \tag{22}$$

then we have

$$\mathbb{P}_{\text{up}}(Q^{-1}(A_\delta)) = \mathbb{P}_{\text{obs}}(A_\delta) = \int_0^\delta 4q \, d\mu_{\mathcal{D}} = 2\delta^2 < \epsilon$$

and the rate of convergence should be of order $O(n^{-1})$ on the rest of the parameter space as desired. Fig. 5 illustrates the numerical recovery of the desired order of convergence on $\Lambda \setminus Q^{-1}(A_\delta)$ with $\epsilon = 0.01$, $\delta = \sqrt{\frac{\epsilon}{2}}$.

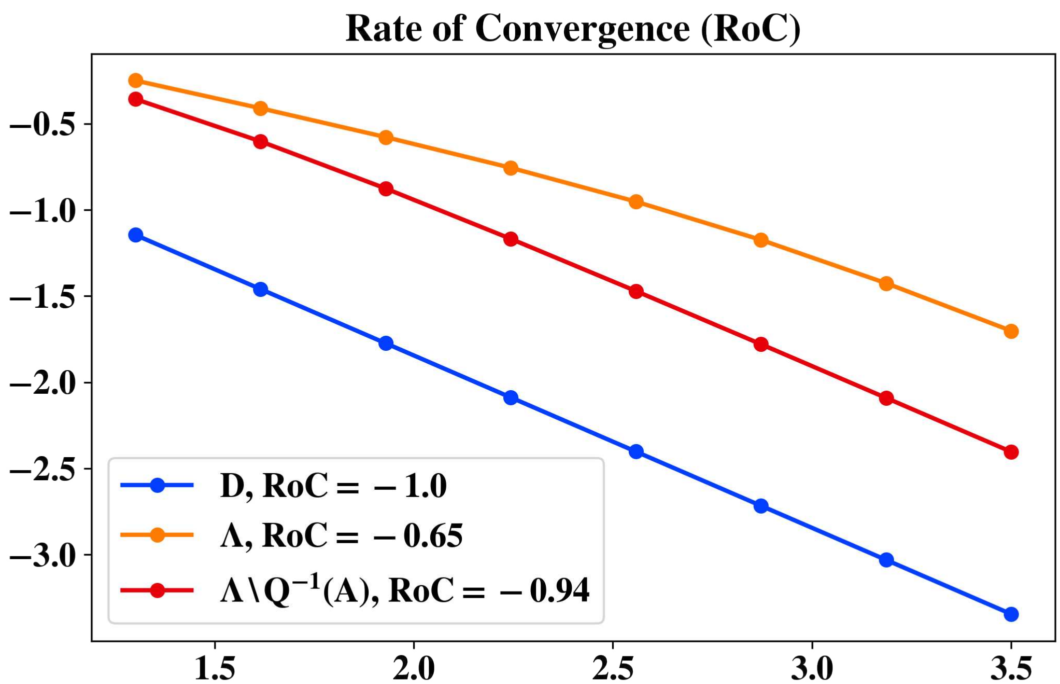

FIG. 5: Shows the rate of convergence (RoC) of $\pi^n_{\text{pred}} \to \pi_{\text{pred}}$ in $L^4(\mathcal{D})$ versus $\pi^n_{\text{up}} \to \pi_{\text{up}}$ in $L^4(\Lambda)$. The rate of convergence of π^n_{up} in $L^4(\Lambda \setminus Q^{-1}(A_\delta))$ is almost $O(n^{-1})$, with A_δ defined by Eq. (22) using $\epsilon = 0.01$ and $\delta = \sqrt{\frac{\epsilon}{2}}$.

7. CONCLUSIONS

This paper addresses the common scenario where finite data or model evaluations are used to approximate probability densities which are subsequently used to construct approximate solutions to stochastic inverse problems. Previous results in the literature demonstrated stability and convergence in the total variation (i.e., the L^1), metric. This paper generalized these results to other methods of quantifying the discrepancy between probability measures that have gained in popularity in recent years, namely, f-divergences, integral probability metrics, and L^p metrics. To the authors knowledge, this paper is the first to theoretically prove and numerically demonstrate stability and convergence for solutions to stochastic inverse problems under these other methods for quantifying discrepancies between measures. Numerical results using straightforward and reproducible test problems illustrated key theoretical results.

ACKNOWLEDGMENTS

This article has been authored by an employee of National Technology & Engineering Solutions of Sandia, LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employee owns all right, title and interest in and to the article and is solely responsible for its contents. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this article or allow others to do so, for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan https://www.energy.gov/downloads/doe-public-access-plan. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

T. Butler and T.Y. Yen's work is supported by the National Science Foundation No. DMS-1818941. T. Butler's work is also supported by the National Science Foundation under Grant No. DMS-2208460 as well as by the NSF IR/D program, while working at National Science Foundation. However, any opinion, finding, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. T.Wildey's work was supported by the DOE Office of Science Early Career Research Program.

REFERENCES

1. A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas. A fast and scalable method for A-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems. *SIAM Journal on Scientific Computing*, 38(1):A243–A272, 2016.
2. S.-I. Amari. *Information Geometry and Its Applications*. Springer Japan, 2016.
3. M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.
4. J. O. Berger, E. Moreno, L. R. Pericchi, M. J. Bayarri, J. M. Bernardo, J. A. Cano, J. De la Horra, J. Martín, D. Ríos-Insúa, B. Betrò, A. Dasgupta, P. Gustafson, L. Wasserman, J. B. Kadane, C. Srinivasan, M. Lavine, A. O'Hagan, W. Polasek, C. P. Robert, C. Goutis, F. Ruggeri, G. Salinetti, and S. Sivaganesan. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.
5. D. Bingham, T. Butler, and D. Estep. Inverse problems for physics-based process models. *Annual Review of Statistics and Its Application*, 11(1), 2024.
6. F. Bisetti, D. Kim, O. Knio, Q. Long, and R. Tempone. Optimal Bayesian experimental design for priors of compact support with application to shock-tube experiments for combustion kinetics. *Int. J. Numerical Methods in Engineering*, 108:136–155, 2016.
7. D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017.
8. V.I. Bogachev. *Measure Theory (Volume 1)*. Springer-Verlag Berlin Heidelberg, 2007.
9. L. Bruder, M.W. Gee, and T. Wildey. Data-consistent solutions to stochastic inverse problems using a probabilistic multi-fidelity method based on conditional densities. *International Journal for Uncertainty Quantificaiton*, 10(5):399–424, 2020.
10. M. Burger and F. Lucka. Maximuma posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators. *Inverse Problems*, 30(11):114004, oct 2014.
11. T. Butler, D. Estep, S. Tavener, C. Dawson, and J.J. Westerink. A measure-theoretic computational method for inverse sensitivity problems iii: Multiple quantities of interest. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):174–202, 2014. doi:10.1137/130930406.
12. T. Butler, J. Jakeman, and T. Wildey. Combining push-forward measures and Bayes' rule to construct consistent solutions to stochastic inverse problems. *SIAM Journal on Scientific Computing*, 40(2):A984–A1011, 2018.
13. T. Butler, J. Jakeman, and T. Wildey. Convergence of Probability Densities Using Approximate Models for Forward

and Inverse Problems in Uncertainty Quantification. *SIAM Journal on Scientific Computing*, 40(5):A3523–A3548, January 2018.

14. T. Butler, J. D. Jakeman, and T. Wildey. Optimal experimental design for prediction based on push-forward probability measures. *Journal of Computational Physics*, 416:109518, 2020.

15. T. Butler, T. Wildey, and T. Y. Yen. Data-consistent inversion for stochastic input-to-output maps. *Inverse problems*, 36(8):85015, 2020.

16. T. Butler, T. Wildey, and W. Zhang. l^p convergence of approximate maps and probability densities for forward and inverse problems in uncertainty quantification. *International Journal for Uncertainty Quantification*, 12(4), 2022.

17. D. Calvetti, J. Kaipio, and E. Somersalo. Inverse problems in the Bayesian framework. *Inverse Problems*, 30(11):110301, 2014.

18. J.T. Change and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51:287–317, 1997.

19. S.L. Cotter, M. Dashti, and A.M. Stuart. Approximation of Bayesian inverse problems. *SIAM Journal of Numerical Analysis*, 48:322–345, 2010.

20. C. Dellacherie and P.A. Meyer. *Probabilities and Potential*. North-Holland Publishing Co., Amsterdam, 1978.

21. D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2), April 1996.

22. B. Fitzpatrick. Bayesian analysis in inverse problems. *Inverse Problems*, 7(5):675, 1991.

23. G.B. Folland. *Real Analysis Modern Techniques and Their Applications, second edition*. John Wiley & Sons, New York, 2008.

24. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, 2013.

25. S. Ghosal and A. van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697 – 723, 2007.

26. A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

27. A. Guntuboyina. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

28. B. E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748, 2008.

29. X. Huan and T.M. Marzouk. Simulation-based optimal bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317, 2013.

30. M. G. Kapteyn, J. V. R. Pretorius, and K. E. Willcox. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science*, 1(5):337–347, May 2021.

31. M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

32. C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos. MMD GAN: Towards deeper understanding of moment matching network. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

33. Qin Li, Li Wang, and Yunan Yang. Differential equation–constrained optimization with stochasticity. *SIAM/ASA Journal on Uncertainty Quantification*, 12(2):549–578, 2024.

34. S. W. Park and J. Kwon. Sphere generative adversarial network based on geometric moment matching. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4287–4296, 2019.

35. N. Petra, J. Martin, G. Stadler, and O. Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems, part ii: Stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM Journal on Scientific Computing*, 36(4):A1525–A1555, 2014.

36. D. Poole and A. E. Raftery. Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255, 2000.

37. Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, 1998.

38. T. Rumbell, J. Parikh, J. Kozloski, and V. Gurev. Novel and flexible parameter estimation methods for data-consistent inversion in mechanistic modeling. *Royal Society Open Science*, 10:230668, 2023.

39. T. Rumbell, C. Wanjiru, I. O. Mulang', S. Obonyo, J. Kozloski, and V. Gurev. Sequential data-consistent model inversion. In *NeurIPS 2023 Workshop on Deep Learning and Inverse Problems*, 2023.

40. Soheil Saghafi, Timothy Rumbell, Viatcheslav Gurev, James Kozloski, Francesco Tamagnini, Kyle C A Wedgwood, and Casey O Diekman. Inferring parameters of pyramidal neuron excitability in mouse models of alzheimer's disease using biophysical modeling and deep learning. *Bulletin of Mathematical Biology*, 86(5):46, March 2024.

41. B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, 1998.

42. S. Sreekumar and Z. Goldfeld. Neural estimation of statistical divergences. *Journal of Machine Learning Research*, 23(126):1–75, 2022.

43. B. K. Sriperumbudur. Mixture density estimation via Hilbert space embedding of measures. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 1027–1030. IEEE, 2011.

44. A. Tran and T. Wildey. Solving stochastic inverse problems for property–structure linkages using data-consistent inversion and machine learning. *The Journal of The Minerals, Metals & Materials Society (TMS)*, 73(1):72–89, 2021.

45. A. Uppal, S. Singh, and B. Poczos. Robust density estimation under Besov IPM losses. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5345–5355. Curran Associates, Inc., 2020.

46. A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK;New York, NY, USA;, 1998.

47. V. Wagner, B. Castellaz, L. Kaiser, S. Höpfl, and N. Radde. Eulerian parameter inference: A probabilistic change of variables for model-based inference with high-variability data sets, 2024. preprint.

48. V. Wagner, S. Höpfl, V. Klingel, M. C. Pop, and N. E. Radde. An inverse transformation algorithm to infer parameter distributions from population snapshot data. *IFAC-PapersOnLine*, 55(23):86–91, 2022. 9th IFAC Conference on Foundations of Systems Biology in Engineering FOSBE 2022.

49. S. N. Walsh, T. M. Wildey, and J. D. Jakeman. Optimal experimental design using a consistent Bayesian approach. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 4(1):011005, 2018.

50. J. Wang, M. Chen, T. Zhao, W. Liao, and Y. Xie. A manifold two-sample test study: integral probability metric with neural networks. *Information and Inference: A Journal of the IMA*, 12(3):1867–1897, 06 2023.

51. J. Wang, M. Wang, and Y. Zhou. Nonlinear wavelet density estimation for biased data in Sobolev spaces. *Journal of Inequalities and Applications*, 2013(1):308, December 2013.

52. C. O. Wu. A cross-validation bandwidth choice for kernel density estimates with selection biased data. *Journal of Multivariate Analysis*, 61(1):38–60, 1997.

53. J. Xu. Convergence rates of wavelet density estimation for negatively dependent sample. *Journal of Inequalities and Applications*, 2019(1):1–13, 2019.

54. T. Y. Yen. *Stochastic Uncertainty Analysis for Data-Consistent Approaches to Inverse Problems*. PhD thesis, University of Colorado Denver, 2021.

55. Y. Zhang and L. Mikelsons. Solving stochastic inverse problems with stochastic BayesFlow. In *2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 966–972, 2023.

APPENDIX A. PROOFS OF TOTAL-VARIATION STABILITY AND CONVERGENCE RESULTS

In this appendix, we provide proofs of the theorems from Section 2, in order of appearance. We begin with Theorem 3.

APPENDIX A.1 Proof of Theorem 3

For fixed measures $\mathbb{P}_{\text{init}}$ and $\mathbb{P}_{\text{obs}}$ with corresponding densities π_{init} and π_{obs} respectively, let $\mathbb{P}^A_{\text{pred}}$ and $\mathbb{P}^B_{\text{pred}}$ denote arbitrary predicted measures which satisfy Assumption 2 with associated updated measures $\mathbb{P}^A_{\text{up}}$ and $\mathbb{P}^B_{\text{up}}$. Additionally, assume there exists another constant $C_1 > 0$ such that

$$\pi_{\text{pred}}(q) \leq C_1 \pi^A_{\text{pred}}(q), \qquad \text{for a.e. } q \in \mathcal{D}.$$

Then, there exists a constant $C_2 > 0$ such that

$$d_{TV}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) \leq C_2 d_{TV}(\mathbb{P}^A_{\text{pred}}, \mathbb{P}^B_{\text{pred}}).$$

Proof. Utilizing (4) for $\mathbb{P}^A_{\text{up}}$ and $\mathbb{P}^B_{\text{up}}$, we obtain

$$d_{TV}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) = \int_\Lambda \left| \pi_{\text{init}}(\lambda) \cdot \frac{\pi_{\text{obs}}(Q(\lambda))}{\pi^A_{\text{pred}}(Q(\lambda))} - \pi_{\text{init}}(\lambda) \cdot \frac{\pi_{\text{obs}}(Q(\lambda))}{\pi^B_{\text{pred}}(Q(\lambda))} \right| d\mu_\Lambda.$$

Collecting and factoring terms, and applying the predictability assumption gives

$$d_{TV}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) \leq C \int_\Lambda \frac{\pi_{\text{init}}(\lambda)}{\pi^A_{\text{pred}}(Q(\lambda))} \cdot \left| \pi^B_{\text{pred}}(Q(\lambda)) - \pi^A_{\text{pred}}(Q(\lambda)) \right| \, d\mu_\Lambda.$$

Applying the disintegration theorem yields

$$d_{TV}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) \leq C \cdot \int_{\mathcal{D}} \int_{\Lambda \cap Q^{-1}(q)} \pi_{\text{init}}(\lambda) \, d\mu_{\Lambda,q} \cdot \frac{1}{\pi^A_{\text{pred}}(q)} \cdot \left| \pi^B_{\text{pred}}(q) - \pi^A_{\text{pred}}(q) \right| \, d\mu_{\mathcal{D}}.$$

Identifying the inner integral as $\pi_{\text{pred}}(q)$ produces

$$d_{TV}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) \leq C \cdot \int_{\mathcal{D}} \frac{\pi_{\text{pred}}(q)}{\pi^A_{\text{pred}}(q)} \cdot \left| \pi^B_{\text{pred}}(q) - \pi^A_{\text{pred}}(q) \right| \, d\mu_{\mathcal{D}},$$

where we now see the utility of the additional assumption on π_{pred} and π^A_{pred}, which allows us to obtain

$$\begin{aligned} d_{TV}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) &\leq C \cdot \int_{\mathcal{D}} \frac{\pi_{\text{pred}}(q)}{\pi^A_{\text{pred}}(q)} \cdot \left| \pi^B_{\text{pred}}(q) - \pi^A_{\text{pred}}(q) \right| \, d\mu_{\mathcal{D}} \\ &\leq C \cdot C_1 \cdot \int_{\mathcal{D}} \left| \pi^B_{\text{pred}}(q) - \pi^A_{\text{pred}}(q) \right| \, d\mu_{\mathcal{D}} \\ &= C_2 \cdot d_{TV}(\mathbb{P}^B_{\text{pred}}, \mathbb{P}^A_{\text{pred}}), \end{aligned}$$

which completes the proof.

□

APPENDIX A.2 Proof of Theorem 4

For fixed measures $\mathbb{P}_{\text{init}}$ and $\mathbb{P}_{\text{pred}}$ with corresponding densities π_{init} and π_{pred} respectively, let $\mathbb{P}^A_{\text{obs}}$ and $\mathbb{P}^B_{\text{obs}}$ denote arbitrary observed measures which satisfy Assumption 2 with associated updated measures $\mathbb{P}^A_{\text{up}}$ and $\mathbb{P}^B_{\text{up}}$. Then,

$$d_{TV}\left(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}\right) = d_{TV}\left(\mathbb{P}^A_{\text{obs}}, \mathbb{P}^B_{\text{obs}}\right).$$

Proof. Utilizing (4) for $\mathbb{P}^A_{\text{up}}$ and $\mathbb{P}^B_{\text{up}}$, we obtain

$$d_{TV}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) = \int_{\Lambda} \left| \pi_{\text{init}}(\lambda) \cdot \frac{\pi^A_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))} - \pi_{\text{init}}(\lambda) \cdot \frac{\pi^B_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))} \right| \, d\mu_{\Lambda}.$$

Factoring out the appropriate terms and applying the disintegration theorem gives

$$d_{TV}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) = \int_{\mathcal{D}} \left(\int_{\Lambda \cap Q^{-1}(q)} \frac{\pi_{\text{init}}(\lambda)}{\pi_{\text{pred}}(Q(\lambda))} \, d\mu_{\Lambda,q} \right) \cdot \left| \pi^A_{\text{obs}}(q) - \pi^B_{\text{obs}}(q) \right| \, d\mu_{\mathcal{D}}.$$

Equation (8) implies the inner integral is one and the conclusion follows. □

APPENDIX B. PROOFS OF F-DIVERGENCE STABILITY AND CONVERGENCE RESULTS

In this appendix, we provide proofs of the theorems from Section 3, in order of appearance. We begin with Theorem 6.

APPENDIX B.1 Proof of Theorem 6: f-divergence and DCI

Given probability measures, $\mathbb{P}_{\text{init}}$, $\mathbb{P}_{\text{obs}}$, and $\mathbb{P}_{\text{pred}}$ which satisfy the Assumption 1 and updated measures $\mathbb{P}_{\text{up}}$ given by (4), we have the following relationship,

$$D_f\left(\mathbb{P}_{\text{up}} \| \mathbb{P}_{\text{init}}\right) = D_f\left(\mathbb{P}_{\text{obs}} \| \mathbb{P}_{\text{pred}}\right).$$

Proof. Utilizing (4) we obtain

$$D_f\left(\mathbb{P}_{\text{up}} \| \mathbb{P}_{\text{init}}\right) = \int_\Lambda f\left(\frac{\pi_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))}\right)\pi_{\text{init}}(\lambda)d\mu_\Lambda = \int_\Lambda f\left(\frac{\pi_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))}\right)d\mathbb{P}_{\text{init}}.$$

Since the predicted measure is the push-forward of the initial, we rewrite this as

$$\int_\Lambda f\left(\frac{\pi_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))}\right)d\mathbb{P}_{\text{init}} = \int_{\mathcal{D}} f\left(\frac{\pi_{\text{obs}}(q)}{\pi_{\text{pred}}(q)}\right)d\mathbb{P}_{\text{pred}}.$$

Substituting $d\mathbb{P}_{\text{pred}} = \pi_{\text{pred}}(q)d\mu_{\mathcal{D}}$ on the right-hand side finishes the proof. □

APPENDIX B.2 Proof of Theorem 7: Stability w.r.t. Predicted with f-divergences

For fixed measures $\mathbb{P}_{\text{init}}$ and $\mathbb{P}_{\text{obs}}$ with corresponding densities π_{init} and π_{obs} respectively, let π_{pred}^A and π_{pred}^B denote predicted densities such that

$$\pi_{\text{obs}}(q) \leq C\pi_{\text{pred}}^A(q), \quad \text{and} \quad \pi_{\text{obs}}(q) \leq C\pi_{\text{pred}}^B(q), \qquad \text{for a.e. } q \in \mathcal{D},$$

for some constant $C > 0$, and let $\mathbb{P}_{\text{up}}^A$ and $\mathbb{P}_{\text{up}}^B$ denotes the respective associated updated measures. Additionally, assume there exists another constant $C_1 > 0$ such that

$$\pi_{\text{pred}}(q) \leq C_1\pi_{\text{pred}}^A(q), \qquad \text{for a.e. } q \in \mathcal{D}.$$

Then, there exists a constant $C_2 > 0$ such that

$$D_f(\mathbb{P}_{\text{up}}^A \,||\, \mathbb{P}_{\text{up}}^B) \leq C_2 \cdot D_f(\mathbb{P}_{\text{pred}}^B \,||\, \mathbb{P}_{\text{pred}}^A).$$

Proof. Utilizing (4) for $\mathbb{P}_{\text{up}}^A$ and $\mathbb{P}_{\text{up}}^B$, we obtain

$$D_f(\mathbb{P}_{\text{up}}^A \,||\, \mathbb{P}_{\text{up}}^B) = \int_\Lambda f\left(\frac{\pi_{\text{init}}(\lambda)\cdot\frac{\pi_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}^A(Q(\lambda))}}{\pi_{\text{init}}(\lambda)\cdot\frac{\pi_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}^B(Q(\lambda))}}\right)\pi_{\text{init}}(\lambda)\cdot\frac{\pi_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}^B(Q(\lambda))}\,d\mu_\Lambda.$$

Canceling terms and applying the predictability assumption gives

$$D_f(\mathbb{P}_{\text{up}}^A \,||\, \mathbb{P}_{\text{up}}^B) \le C\int_\Lambda \pi_{\text{init}}(\lambda)\cdot f\left(\frac{\pi_{\text{pred}}^B(Q(\lambda))}{\pi_{\text{pred}}^A(Q(\lambda))}\right)\,d\mu_\Lambda.$$

Applying the disintegration theorem yields

$$D_f(\mathbb{P}_{\text{up}}^A \,||\, \mathbb{P}_{\text{up}}^B) \le C\cdot\int_{\mathcal{D}}\int_{\Lambda\cap Q^{-1}(q)} \pi_{\text{init}}(\lambda)\,d\mu_{\Lambda,q}\cdot f\left(\frac{\pi_{\text{pred}}^B(q)}{\pi_{\text{pred}}^A(q)}\right)\,d\mu_{\mathcal{D}}.$$

Identifying the inner integral as $\pi_{\text{pred}}(q)$ produces

$$D_f(\mathbb{P}_{\text{up}}^A \,||\, \mathbb{P}_{\text{up}}^B) \le C\cdot\int_{\mathcal{D}} f\left(\frac{\pi_{\text{pred}}^B(q)}{\pi_{\text{pred}}^A(q)}\right)\pi_{\text{pred}}(q)\,d\mu_{\mathcal{D}},$$

where we now see the utility of the additional assumption on π_{pred} and π_{pred}^A, which allows us to obtain

$$\begin{aligned}
D_f(\mathbb{P}_{\text{up}}^A \,||\, \mathbb{P}_{\text{up}}^B) &\le C\cdot\int_{\mathcal{D}} f\left(\frac{\pi_{\text{pred}}^B(q)}{\pi_{\text{pred}}^A(q)}\right)\pi_{\text{pred}}(q)\,d\mu_{\mathcal{D}} \\
&\le C\cdot C_1\cdot\int_{\mathcal{D}} f\left(\frac{\pi_{\text{pred}}^B(q)}{\pi_{\text{pred}}^A(q)}\right)\pi_{\text{pred}}^A(q)\,d\mu_{\mathcal{D}} \\
&= C_2\cdot D_f(\mathbb{P}_{\text{pred}}^B \,||\, \mathbb{P}_{\text{pred}}^A),
\end{aligned}$$

which completes the proof. □

APPENDIX B.3 Proof of Theorem 9: Stability w.r.t. Observed with f-divergences

For fixed measures $\mathbb{P}_{\text{init}}$ and $\mathbb{P}_{\text{pred}}$ with corresponding densities π_{init} and π_{pred} respectively, let $\mathbb{P}_{\text{obs}}^A$ and $\mathbb{P}_{\text{obs}}^B$ denote observed measures such that

$$\pi_{\text{obs}}^A(q) \le C\pi_{\text{pred}}(q), \quad \text{and} \quad \pi_{\text{obs}}^B(q) \le C\pi_{\text{pred}}(q) \quad \text{for a.e. } q\in\mathcal{D},$$

for some constant $C > 0$, and let $\mathbb{P}^A_{\text{up}}$ and $\mathbb{P}^B_{\text{up}}$ denote the respective associated updated measures. Then,

$$D_f(\mathbb{P}^A_{\text{up}} \,||\, \mathbb{P}^B_{\text{up}}) = D_f(\mathbb{P}^A_{\text{obs}} \,||\, \mathbb{P}^B_{\text{obs}}).$$

Proof. Utilizing (4) for $\mathbb{P}^A_{\text{up}}$ and $\mathbb{P}^B_{\text{up}}$, we obtain

$$D_f(\mathbb{P}^A_{\text{up}} \,||\, \mathbb{P}^B_{\text{up}}) = \int_\Lambda f\left(\frac{\pi_{\text{init}}(\lambda)\cdot \frac{\pi^A_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))}}{\pi_{\text{init}}(\lambda)\cdot \frac{\pi^B_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))}}\right)\pi_{\text{init}}(\lambda)\cdot\frac{\pi^B_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))}\,d\mu_\Lambda.$$

Canceling terms and applying the disintegration theorem gives

$$D_f(\mathbb{P}^A_{\text{up}} \,||\, \mathbb{P}^B_{\text{up}}) = \int_\mathcal{D}\left(\int_{\Lambda\cap Q^{-1}(q)}\frac{\pi_{\text{init}}(\lambda)}{\pi_{\text{pred}}(Q(\lambda))}\,d\mu_{\Lambda,q}\right)\cdot f\left(\frac{\pi^A_{\text{obs}}(q)}{\pi^B_{\text{obs}}(q)}\right)\pi^B_{\text{obs}}(q)\,d\mu_\mathcal{D}.$$

Equation (8) implies the inner integral is one and the conclusion follows. □

APPENDIX C. TOTAL VARIATION AS IPM

We show that choosing $\mathcal{F}$ to be $\{f : ||f||_\infty \leq 1\}$ produces the total variation metric. Consider,

$$\begin{aligned} d_\mathcal{F}(\mathbb{P}^A, \mathbb{P}^B) &= \sup_{f\in\mathcal{F}}\left|\int_\mathcal{X} f d\mathbb{P}^A - \int_\mathcal{X} f d\mathbb{P}^B\right| \\ &= \sup_{f\in\mathcal{F}}\left|\int_\mathcal{X} f(x)\pi^A(x)d\mu_\mathcal{X} - \int_\mathcal{X} f(x)\pi^B(x)d\mu_\mathcal{X}\right| \\ &= \sup_{f\in\mathcal{F}}\left|\int_\mathcal{X} f(x)(\pi^A(x)-\pi^B(x))d\mu_\mathcal{X}\right|. \end{aligned}$$

Now it is clear that for every $f \in \mathcal{F}$ where $\mathcal{F}$ is chosen to be $\{f : ||f||_\infty \leq 1\}$,

$$\begin{aligned} \left|\int_\mathcal{X} f(x)(\pi^A(x)-\pi^B(x))d\mu_\mathcal{X}\right| &\leq \int_\mathcal{X} |f(x)|\left|\pi^A(x)-\pi^B(x)\right| d\mu_\mathcal{X} \\ &\leq \int_\mathcal{X} ||f||_\infty \left|\pi^A(x)-\pi^B(x)\right| d\mu_\mathcal{X} \\ &\leq \int_\mathcal{X} \left|\pi^A(x)-\pi^B(x)\right| d\mu_\mathcal{X} \\ &= d_{TV}(\mathbb{P}^A, \mathbb{P}^B) \end{aligned}$$

by the triangle inequality and the fact that $||f||_\infty \leq 1$. For the other direction, define $f_\pm(x)$ to be

$$f_\pm(x) = \begin{cases} 1 & \pi^A(x) - \pi^B(x) > 0 \\ -1 & \pi^A(x) - \pi^B(x) < 0. \end{cases}$$

Then $||f_\pm||_\infty \leq 1$. In addition, $\forall x \in \mathcal{X}$, the definition of $f_\pm$ implies that

$$\begin{aligned} \left|\pi^A(x) - \pi^B(x)\right| &\leq f_\pm(x)(\pi^A(x) - \pi^B(x)) \\ \Rightarrow \int_\mathcal{X} \left|\pi^A(x) - \pi^B(x)\right| d\mu_\mathcal{X} &\leq \int_\mathcal{X} f_\pm(x)(\pi^A(x) - \pi^B(x)) d\mu_\mathcal{X} \\ \Rightarrow d_{TV}(\mathbb{P}^A, \mathbb{P}^B) &\leq \left|\int_\mathcal{X} f_\pm(x)(\pi^A(x) - \pi^B(x)) d\mu_\mathcal{X}\right| \\ &\leq \sup_{f\in\mathcal{F}} \left|\int_\mathcal{X} f(x)(\pi^A(x) - \pi^B(x)) d\mu_\mathcal{X}\right| \\ &= d_\mathcal{F}(\mathbb{P}^A, \mathbb{P}^B) \end{aligned}$$

Thus, the total variation metric is equivalent to the integral probability metric with $\mathcal{F}$ chosen to be $\{f : ||f||_\infty \leq 1\}$ since

$$d_\mathcal{F}(\mathbb{P}^A, \mathbb{P}^B) \leq d_{TV}(\mathbb{P}^A, \mathbb{P}^B) \quad \text{and } d_{TV}(\mathbb{P}^A, \mathbb{P}^B) \leq d_\mathcal{F}(\mathbb{P}^A, \mathbb{P}^B).$$

APPENDIX D. PROOFS OF IPM STABILITY AND CONVERGENCE RESULTS

In this appendix, we provide proofs of the theorems from Section 4, in order of appearance. We begin with Theorem 11.

APPENDIX D.1 Proof of Theorem 11: Stability of Updated via Predicted using IPM

Let $\mathcal{F}_\Lambda$ and $\mathcal{G}_\mathcal{D}$ be used to define IPM for measures on Λ and $\mathcal{D}$, respectively. Suppose $\mathbb{E}_{\Lambda|q}$ is a bounded operator from $\mathcal{F}_\Lambda$ to $\mathcal{G}_\mathcal{D}$. For fixed measures $\mathbb{P}_{\text{init}}$ and $\mathbb{P}_{\text{obs}}$ with corresponding densities π_{init} and π_{obs} respectively, let π^A_{pred} and π^B_{pred} denote predicted densities satisfying Assumption 2.2 and let $\mathbb{P}^A_{\text{up}}$ and $\mathbb{P}^B_{\text{up}}$ denotes the respective associated updated measures. Additionally, assume there exists another constant $C_1 > 0$ such that

$$\pi_{\text{pred}}(q) \leq C_1 \pi^A_{\text{pred}}(q), \quad \text{for a.e. } q \in \mathcal{D}.$$

Then, there exists a constant $C_2 > 0$ such that

$$d_{\mathcal{F}_\Lambda}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) \leq C_2 d_{\mathcal{G}_\mathcal{D}}(\mathbb{P}^A_{\text{pred}}, \mathbb{P}^B_{\text{pred}})$$

Proof. Without loss of generality, assume that $\mathcal{F}_\Lambda := \{f : ||f||_{\mathcal{F}_\Lambda} \leq 1\}$ and $\mathcal{G}_\mathcal{D} := \{g : ||g||_{\mathcal{G}_\mathcal{D}} \leq 1\}$, where $||\cdot||_{\mathcal{F}_\Lambda}$ and $||\cdot||_{\mathcal{G}_\mathcal{D}}$ denote the norms defining the IPM for measures on Λ and $\mathcal{D}$, respectively. Since $\mathbb{E}_{\Lambda|q}$ is a bounded operator, $\exists C > 0$ such that $\forall f \in \mathcal{F}_\Lambda$

$$||\mathbb{E}_{\Lambda|q}(f)||_{\mathcal{G}_\mathcal{D}} \leq C||f||_{\mathcal{F}_\Lambda} \leq C.$$

Thus,

$$\frac{1}{C}||\mathbb{E}_{\Lambda|q}(f)||_{\mathcal{G}_\mathcal{D}} \leq 1 \Rightarrow \frac{1}{C}\mathbb{E}_{\Lambda|q}(f) \in \mathcal{G}_\mathcal{D}$$

In other words, the range of $\frac{1}{C}\mathbb{E}_{\Lambda|q}$ is contained in $\mathcal{G}_\mathcal{D}$.

Therefore, we have,

$$\begin{aligned}
d_{\mathcal{F}_\Lambda}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) &= \sup_{f\in\mathcal{F}_\Lambda} \left| \int_\mathcal{D} \int_{\Lambda\cap Q^{-1}(q)} f(\lambda)\pi_{\text{init}\,|q}(\lambda)\, d\mu_{\Lambda,q}\pi_{\text{pred}}(q) \left(\frac{\pi_{\text{obs}}(q)}{\pi^A_{\text{pred}}(q)} - \frac{\pi_{\text{obs}}(q)}{\pi^B_{\text{pred}}(q)} \right) d\mu_\mathcal{D} \right| \\
&= \sup_{f\in\mathcal{F}_\Lambda} \left| \int_\mathcal{D} \mathbb{E}_{\Lambda|q}(f) \frac{\pi_{\text{obs}}(q)\pi_{\text{pred}}(q)}{\pi^A_{\text{pred}}(q)\pi^B_{\text{pred}}(q)} (\pi^B_{\text{pred}}(q) - \pi^A_{\text{pred}}(q))\, d\mu_\mathcal{D} \right| \\
&= C \cdot \sup_{f\in\mathcal{F}_\Lambda} \left| \int_\mathcal{D} \frac{1}{C}\mathbb{E}_{\Lambda|q}(f) \frac{\pi_{\text{obs}}(q)\pi_{\text{pred}}(q)}{\pi^A_{\text{pred}}(q)\pi^B_{\text{pred}}(q)} (\pi^B_{\text{pred}}(q) - \pi^A_{\text{pred}}(q))\, d\mu_\mathcal{D} \right| \\
&\leq C \cdot \sup_{g\in\mathcal{G}_\mathcal{D}} \left| \int_\mathcal{D} g(q) \frac{\pi_{\text{obs}}(q)\pi_{\text{pred}}(q)}{\pi^A_{\text{pred}}(q)\pi^B_{\text{pred}}(q)} (\pi^B_{\text{pred}}(q) - \pi^A_{\text{pred}}(q))\, d\mu_\mathcal{D} \right| \\
&\leq C_2 \cdot d_{\mathcal{G}_\mathcal{D}}(\mathbb{P}^A_{\text{pred}}, \mathbb{P}^B_{\text{pred}}),
\end{aligned}$$

where we have used the predictability assumption and the additional assumption involving π_{pred} and π^A_{pred} to obtain the last inequality.

□

APPENDIX D.2 Proof of Theorem 12: Stability of Updated via Observed using IPM

Let $\mathcal{F}_\Lambda$ and $\mathcal{G}_\mathcal{D}$ be used to define IPM for measures on Λ and $\mathcal{D}$, respectively. Suppose $\mathbb{E}_{\Lambda|q}$ is a bounded operator from $\mathcal{F}_\Lambda$ to $\mathcal{G}_\mathcal{D}$. For fixed measures $\mathbb{P}_{\text{init}}$ and $\mathbb{P}_{\text{pred}}$ with corresponding densities π_{init} and π_{pred} respectively, let $\mathbb{P}^A_{\text{obs}}$ and $\mathbb{P}^B_{\text{obs}}$ denote observed measures satisfying Assumption 2.1 and let $\mathbb{P}^A_{\text{up}}$ and $\mathbb{P}^B_{\text{up}}$ denote the respective associated updated measures. Then, there exists $C > 0$ such that

$$d_{\mathcal{F}_\Lambda}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) \leq C d_{\mathcal{G}_\mathcal{D}}(\mathbb{P}^A_{\text{obs}}, \mathbb{P}^B_{\text{obs}})$$

Proof. For the case of approximate predicted densities, the proof is similar to the proof in APPENDIX D.1, except we have

$$\begin{aligned}
d_{\mathcal{F}_\Lambda}(\mathbb{P}^A_{\text{up}}, \mathbb{P}^B_{\text{up}}) &= \sup_{f \in \mathcal{F}_\Lambda} \left| \int_\mathcal{D} \int_{\Lambda \cap Q^{-1}(q)} f(\lambda) \pi_{\text{init}\,|q}(\lambda)\, d\mu_{\Lambda,q} (\pi^A_{\text{obs}}(q) - \pi^B_{\text{obs}}(q))\, d\mu_\mathcal{D} \right| \\
&= \sup_{f \in \mathcal{F}_\Lambda} \left| \int_\mathcal{D} \mathbb{E}_{\Lambda|q}(f)(\pi^A_{\text{obs}}(q) - \pi^B_{\text{obs}}(q))\, d\mu_\mathcal{D} \right| \\
&= C \cdot \sup_{f \in \mathcal{F}_\Lambda} \left| \int_\mathcal{D} \frac{1}{C} \mathbb{E}_{\Lambda|q}(f)(\pi^A_{\text{obs}}(q) - \pi^B_{\text{obs}}(q))\, d\mu_\mathcal{D} \right| \\
&\leq C \cdot \sup_{g \in \mathcal{G}_\mathcal{D}} \left| \int_\mathcal{D} g(q)(\pi^A_{\text{obs}}(q) - \pi^B_{\text{obs}}(q))\, d\mu_\mathcal{D} \right| \\
&= C \cdot d_{\mathcal{G}_\mathcal{D}}(\mathbb{P}^A_{\text{obs}}, \mathbb{P}^B_{\text{obs}})
\end{aligned}$$

□

APPENDIX D.3 Proof of Theorem 14: Stability using the Pullback IPM

For fixed measures $\mathbb{P}_{\text{init}}$ and $\mathbb{P}_{\text{pred}}$ with corresponding densities π_{init} and π_{pred} respectively, let $\mathbb{P}^A_{\text{obs}}$ and $\mathbb{P}^B_{\text{obs}}$ denote observed measures such that

$$\pi^A_{\text{obs}}(q) \leq C\pi_{\text{pred}}(q), \quad \text{and} \quad \pi^B_{\text{obs}}(q) \leq C\pi_{\text{pred}}(q) \quad \text{for a.e. } q \in \mathcal{D},$$

for some constant $C > 0$ (i.e., Assumption 2.1), and let $\mathbb{P}^A_{\text{up}}$ and $\mathbb{P}^B_{\text{up}}$ denote the respective associated updated measures. Given an IPM on $\mathcal{D}$ defined by $\mathcal{F}_\mathcal{D}$ and the corresponding data-consistent IPM defined by $\mathcal{F}_\Lambda^*$, we have

$$d_{\mathcal{F}_\Lambda^*}(\mathbb{P}_{\text{up}}^A, \mathbb{P}_{\text{up}}^B) = d_{\mathcal{F}_\mathcal{D}}(\mathbb{P}_{\text{obs}}^A, \mathbb{P}_{\text{obs}}^B), \tag{D.1}$$

Similarly, for fixed measures $\mathbb{P}_{\text{init}}$ and $\mathbb{P}_{\text{obs}}$ with corresponding densities π_{init} and π_{obs} respectively, let π_{pred}^A and π_{pred}^B denote predicted densities such that

$$\pi_{\text{obs}}(q) \leq C\pi_{\text{pred}}^A(q), \quad \text{and} \quad \pi_{\text{obs}}(q) \leq C\pi_{\text{pred}}^B(q), \quad \text{for a.e. } q \in \mathcal{D},$$

for some constant $C > 0$ (i.e., Assumption 2.2), and let $\mathbb{P}_{\text{up}}^A$ and $\mathbb{P}_{\text{up}}^B$ denote the respective associated updated measures. Additionally, assume there exists another constant $C_1 > 0$ such that

$$\pi_{\text{pred}}(q) \leq C_1\pi_{\text{pred}}^A(q), \quad \text{for a.e. } q \in \mathcal{D}.$$

Then, there exists a constant $C_2 > 0$ such that

$$d_{\mathcal{F}_\Lambda^*}(\mathbb{P}_{\text{up}}^A, \mathbb{P}_{\text{up}}^B) \leq C_2 d_{\mathcal{F}_\mathcal{D}}(\mathbb{P}_{\text{pred}}^A, \mathbb{P}_{\text{pred}}^B). \tag{D.2}$$

Proof. Given $f \in \mathcal{F}_\Lambda^*$, choose $g_f \in \mathcal{F}_\mathcal{D}$ to be the corresponding function such that

$$f(\lambda) = g_f(Q(\lambda))$$

which exists because of the definition of the data-consistent IPM.

$$\begin{aligned}
\mathbb{E}_{\Lambda|q}(f) &= \int_{\Lambda\cap Q^{-1}(q)} f(\lambda)\pi_{\text{init}\,|q}(\lambda)d\mu_{\Lambda,q} \\
&= \int_{\Lambda\cap Q^{-1}(q)} g_f(Q(\lambda))\pi_{\text{init}\,|q}(\lambda)d\mu_{\Lambda,q} \\
&= g(q)\int_{\Lambda\cap Q^{-1}(q)} \pi_{\text{init}\,|q}(\lambda)d\mu_{\Lambda,q} \\
&= g(q).
\end{aligned}$$

Similarly, for each $g \in \mathcal{F}_\mathcal{D}$, choose f_g to be the corresponding function in $\mathcal{F}_\Lambda^*$. The same equality holds for each function g. Thus, to prove (11), we have

$$\begin{aligned} d_{\mathcal{F}_\Lambda^*}(\mathbb{P}_{\text{up}}^A, \mathbb{P}_{\text{up}}^B) &= \sup_{f \in \mathcal{F}_\Lambda^*} \left| \int_{\mathcal{D}} \mathbb{E}_{\Lambda|q}(f) \left(\frac{\pi_{\text{obs}}(q)}{\pi_{\text{pred}}^B(q)} \right) \left(\frac{\pi_{\text{pred}}(q)}{\pi_{\text{pred}}^A(q)} \right) (\pi_{\text{pred}}^B(q) - \pi_{\text{pred}}^A(q)) \, d\mu_{\mathcal{D}} \right| \\ &\leq C \cdot C_1 \sup_{g \in \mathcal{G}_{\mathcal{D}}} \left| \int_{\mathcal{D}} g(q)(\pi_{\text{pred}}^A(q) - \pi_{\text{pred}}^B(q)) \, d\mu_{\mathcal{D}} \right| \\ &= d_{\mathcal{G}_{\mathcal{D}}}(\mathbb{P}_{\text{pred}}^A, \mathbb{P}_{\text{pred}}^B). \end{aligned}$$

To prove (12), we proceed as above except we do not require the additional assumption,

$$\begin{aligned} d_{\mathcal{F}_\Lambda^*}(\mathbb{P}_{\text{up}}^A, \mathbb{P}_{\text{up}}^B) &= \sup_{f \in \mathcal{F}_\Lambda^*} \left| \int_{\mathcal{D}} \mathbb{E}_{\Lambda|q}(f)(\pi_{\text{obs}}^A(q) - \pi_{\text{obs}}^B(q)) \, d\mu_{\mathcal{D}} \right| \\ &= \sup_{g \in \mathcal{G}_{\mathcal{D}}} \left| \int_{\mathcal{D}} g(q)(\pi_{\text{obs}}^A(q) - \pi_{\text{obs}}^B(q)) \, d\mu_{\mathcal{D}} \right| \\ &= d_{\mathcal{G}_{\mathcal{D}}}(\mathbb{P}_{\text{obs}}, \widetilde{\mathbb{P}}_{\text{obs}}). \end{aligned}$$

□

APPENDIX E. PROOFS OF L^P CONVERGENCE RESULTS

In this appendix, we provide proofs of the theorems from Section 5, in order of appearance. We begin with Theorem 15.

APPENDIX E.1 Proof of Theorem 15: L^p Convergence with Approximated Predicted Densities

Suppose $\pi_{\text{init}} \in L^\infty(\Lambda)$ and π_{obs} are chosen so that Assumption 1 is satisfied. If (π_{pred}^n) satisfies Assumption 3 and $\pi_{\text{pred}}^n \to \pi_{\text{pred}}$ in $L^p(\mathcal{D})$, then $\pi_{\text{up}}^n \to \pi_{\text{up}}$ in $L^p(\Lambda)$.

The proof uses standard measure-theoretic techniques. First, we partition the output space into two sets of "small" and "large" measure. We then consider the pre-images of these sets in the input space and separately argue why the L^p difference between the approximate and exact updated densities are small on each of these sets. The argument for the pre-image of the "small" set is straightforward in that it directly relies upon the fact that the initial probability of the set is itself small. The argument for the pre-image of the "large" set is more subtle.

Proof. Let $\epsilon > 0$. Since π_{pred} is a probability density and therefore in $L^1(\mathcal{D})$, we can choose a set $A_\delta \subset \mathcal{D}$

defined by $\delta > 0$ as

$$A_\delta := \{q : \pi_{\text{pred}}(q) < \delta\}$$

such that

$$\int_{A_\delta} \pi_{\text{pred}}(q)\, d\mu_{\mathcal{D}} < \frac{\epsilon^p}{3 \cdot 2^{p-1}} \cdot \frac{1}{C^p \cdot ||\pi_{\text{init}}(\lambda)||^{p-1}_{L^\infty(\Lambda)}},$$

where C is the maximum of the predictability constants from Assumptions 1 and 3. We use the set $Q^{-1}(A_\delta) \subset \Lambda$ to split the following integral into two terms that we can separately bound,

$$\begin{aligned} ||\pi^n_{\text{up}} - \pi_{\text{up}}||^p_{L^p(\Lambda)} &= \int_\Lambda \left|\pi^n_{\text{up}}(\lambda) - \pi_{\text{up}}(\lambda)\right|^p \, d\mu_\Lambda \\ &= \underbrace{\int_{\Lambda \setminus Q^{-1}(A_\delta)} \left|\pi^n_{\text{up}}(\lambda) - \pi_{\text{up}}(\lambda)\right|^p \, d\mu_\Lambda}_{=:I_{\Lambda \setminus Q^{-1}(A_\delta)}} \\ &\quad + \underbrace{\int_{Q^{-1}(A_\delta)} \left|\pi^n_{\text{up}}(\lambda) - \pi_{\text{up}}(\lambda)\right|^p \, d\mu_\Lambda}_{=:I_{Q^{-1}(A_\delta)}}. \end{aligned}$$

First, consider the "small" set $Q^{-1}(A_\delta)$. We rewrite the approximate updated density and true updated density in terms of the initial density times the ratio $r_n(q) = \frac{\pi_{\text{obs}}(q)}{\pi^n_{\text{pred}}(q)}$ and $r(q) = \frac{\pi_{\text{obs}}(q)}{\pi_{\text{pred}}(q)}$ respectively. From Assumptions 1 and 3, there exists N_C such that $\forall n \geq N_C$

$$r_n(q) = \frac{\pi_{\text{obs}}(q)}{\pi^n_{\text{pred}}(q)} \leq C \quad \text{and} \quad r(q) = \frac{\pi_{\text{obs}}(q)}{\pi_{\text{pred}}(q)} \leq C.$$

Thus,

$$\begin{aligned} I_{Q^{-1}(A_\delta)} &= \int_{Q^{-1}(A_\delta)} |\pi_{\text{init}}(\lambda) r_n(Q(\lambda)) - \pi_{\text{init}}(\lambda) r(Q(\lambda))|^p \, d\mu_\Lambda \\ &= \int_{Q^{-1}(A_\delta)} |\pi_{\text{init}}(\lambda)|^p \, |r_n(Q(\lambda)) - r(Q(\lambda))|^p \, d\mu_\Lambda \\ &\leq 2^p C^p \cdot \int_{Q^{-1}(A_\delta)} |\pi_{\text{init}}(\lambda)|^p \, d\mu_\Lambda. \end{aligned}$$

Applying Hölder's inequality $p-1$ times followed by the disintegration theorem gives

$$\begin{aligned}\int_{Q^{-1}(A_\delta)} |\pi_{\text{init}}(\lambda)|^p \, d\mu_\Lambda &\le ||\pi_{\text{init}}(\lambda)||^{p-1}_{L^\infty(\Lambda)} \cdot \int_{Q^{-1}(A_\delta)} \pi_{\text{init}}(q)\, d\mu_\Lambda, \\ &= ||\pi_{\text{init}}(\lambda)||^{p-1}_{L^\infty(\Lambda)} \cdot \int_{A_\delta} \underbrace{\int_{\Lambda\cap Q^{-1}(q)} \pi_{\text{init}}(\lambda)\, d\mu_{\Lambda,q}}_{=\pi_{\text{pred}}(q)}\, d\mu_{\mathcal{D}} \\ &= ||\pi_{\text{init}}(\lambda)||^{p-1}_{L^\infty(\Lambda)} \cdot \int_{A_\delta} \pi_{\text{pred}}(q)\, d\mu_{\mathcal{D}}.\end{aligned}$$

By our choice of A_δ,

$$I_{Q^{-1}(A_{\delta_\epsilon})} \le \frac{2\epsilon^p}{3}.$$

Next, we bound the integral on the "large" set $\Lambda \setminus Q^{-1}(A_\delta)$. We begin by re-arranging the terms of the difference between updated densities by finding a common denominator as follows

$$\begin{aligned}\left|\pi^n_{\text{up}}(\lambda) - \pi_{\text{up}}(\lambda)\right| &= \pi_{\text{init}}(\lambda)\left|\frac{\pi_{\text{obs}}(Q(\lambda))}{\pi^n_{\text{pred}}(Q(\lambda))} - \frac{\pi_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))}\right| \\ &= \pi_{\text{init}}(\lambda)\cdot\pi_{\text{obs}}(Q(\lambda))\cdot\left|\frac{\pi^n_{\text{pred}}(Q(\lambda)) - \pi_{\text{pred}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))\widetilde{\pi}^n_{\text{pred}}(Q(\lambda))}\right| \\ &= \frac{\pi_{\text{init}}(\lambda)\cdot\pi_{\text{obs}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))\cdot\pi^n_{\text{pred}}(Q(\lambda))}\cdot\left|\pi^n_{\text{pred}}(Q(\lambda)) - \pi_{\text{pred}}(Q(\lambda))\right| \\ &= \frac{\pi_{\text{init}}(\lambda)}{\pi_{\text{pred}}(Q(\lambda))}\cdot r_n(Q(\lambda))\cdot\left|\pi^n_{\text{pred}}(Q(\lambda)) - \pi_{\text{pred}}(Q(\lambda))\right|\end{aligned}$$

where $r_n(q)$ is the ratio described earlier. Assumption 3 implies $r_n(Q(\lambda))$ is bounded by C and $\pi_{\text{pred}}(q) \ge \delta$ on the complement of A_δ. It follows that

$$I_{\Lambda\setminus Q^{-1}(A_\delta)} \le \frac{C^p}{\delta^{p-1}} \int_{\Lambda\setminus Q^{-1}(A_\delta)} |\pi_{\text{init}}(\lambda)|^p \cdot \frac{\left|\pi^n_{\text{pred}}(Q(\lambda)) - \pi_{\text{pred}}(Q(\lambda))\right|^p}{\pi_{\text{pred}}(Q(\lambda))} d\mu_\Lambda.$$

Rewriting the above integrand as

$$|\pi_{\text{init}}(\lambda)|^{p-1}\cdot\frac{\pi_{\text{init}}(Q(\lambda))}{\pi_{\text{pred}}(Q(\lambda))}\left|\pi^n_{\text{pred}}(Q(\lambda)) - \pi_{\text{pred}}(Q(\lambda))\right|^p,$$

and then applying Hölder's inequality $p-1$ times, we obtain

$$I_{\Lambda\setminus Q^{-1}(A_\delta)} \leq \frac{C^p ||\pi_{\text{init}}(\lambda)||^{p-1}_{L^\infty(\Lambda)}}{\delta^{p-1}} \int_{\Lambda\setminus Q^{-1}(A_\delta)} \frac{\pi_{\text{init}}(\lambda)}{\pi_{\text{pred}}(Q(\lambda))} \left|\pi^n_{\text{pred}}(Q(\lambda)) - \pi_{\text{pred}}(Q(\lambda))\right|^p d\mu_\Lambda.$$

Applying the disintegration theorem yields

$$I_{\Lambda\setminus Q^{-1}(A_\delta)} \leq \frac{C^p ||\pi_{\text{init}}(\lambda)||^{p-1}_{L^\infty(\Lambda)}}{\delta^{p-1}} \int_{\mathcal{D}} \underbrace{\int_{\Lambda\setminus Q^{-1}(A_\delta)} \frac{\pi_{\text{init}}(\lambda)}{\pi_{\text{pred}}(Q(\lambda))} d\mu_{\Lambda,q}}_{=1} \left|\pi^n_{\text{pred}}(q) - \pi_{\text{pred}}(q)\right|^p d\mu_{\mathcal{D}},$$

which reduces to

$$\begin{aligned} I_{\Lambda\setminus Q^{-1}(A_\delta)} &\leq \frac{C^p ||\pi_{\text{init}}(\lambda)||^{p-1}_{L^\infty(\Lambda)}}{\delta^{p-1}} \int_{\mathcal{D}} \left|\pi^n_{\text{pred}}(q) - \pi_{\text{pred}}(q)\right|^p d\mu_{\mathcal{D}} \\ &= \frac{C^p ||\pi_{\text{init}}(\lambda)||^{p-1}_{L^\infty(\Lambda)}}{\delta^{p-1}} ||\pi^n_{\text{pred}} - \pi_{\text{pred}}||^p_{L^p(\mathcal{D})}. \end{aligned} \tag{E.1}$$

Since $\pi^n_{\text{pred}} \to \pi_{\text{pred}}$ in $L^p(\mathcal{D})$, we can choose $N_\delta \geq N_C$ such that $n \geq N_\delta$ implies that the above integral is less than $\epsilon^p/3$. Combining this with the bound from the "small" set, we have that for $n \geq N_\delta$,

$$\begin{aligned} ||\pi^n_{\text{up}} - \pi_{\text{up}}||_{L^p(\Lambda)} &\leq \left(I_{Q^{-1}(A_\delta)} + I_{\Lambda\setminus Q^{-1}(A_\delta)}\right)^{1/p} \\ &< \left(\frac{2\epsilon^p}{3} + \frac{\epsilon^p}{3}\right)^{1/p} = \epsilon. \end{aligned}$$

The conclusion follows. □

APPENDIX E.2 Proof of Theorem 16: Rate of Convergence with Predicted in L_p

Suppose $\pi_{\text{init}} \in L^\infty(\Lambda)$ and π_{obs} are chosen so that Assumption 1 is satisfied. If (π^n_{pred}) satisfies Assumption 3, $\pi^n_{\text{pred}} \to \pi_{\text{pred}}$ in $L^p(\mathcal{D})$, and the convergence rate of $\mathbb{P}^n_{\text{pred}}$ is of order $O(\rho(n))$ on almost all of $\mathcal{D}$, then the convergence rate of $\mathbb{P}^n_{\text{up}}$ is of order $O(\rho(n))$ on almost all of Λ.

Proof. This follows immediately from the bound obtained in Equation (E.1) in the proof of Theorem 15 located in Appendix APPENDIX E.1. □

APPENDIX E.3 Proof of Theorem 17: L^p Convergence with Approximated Observed Densities

Suppose $\pi_{\text{init}} \in L^\infty(\Lambda)$ and π_{obs} are chosen so that Assumption 1 is satisfied. If (π^n_{obs}) satisfies Assumption 3 and $\pi^n_{\text{obs}} \to \pi_{\text{obs}}$ in $L^p(\mathcal{D})$, then $\pi^n_{\text{up}} \to \pi_{\text{up}}$ in $L^p(\Lambda)$.

The proof is similar to that of Theorem 15 in that we let $\epsilon > 0$ and follow analogous (and in some case identical) steps to choose an N such that $n \geq N$ implies that $||\widetilde{\pi}^n_{\text{up}} - \pi_{\text{up}}||_{L^p(\Lambda)} < \epsilon$. Below, we mention the relevant, and in some cases subtle, details that change in the argument.

Proof. In proving that $I_{Q^{-1}(A_\delta)}$ is small, the only relevant detail that changes is that $r_n(q)$ is now defined in terms of the ratio of the approximated observed density π^n_{obs} to π_{pred}. The proof that $I_{\Lambda\setminus Q^{-1}(A_\delta)}$ can be made small for sufficiently large n is simpler than in the previous proof. First, there is no need to find a common denominator in the difference of the approximated and exact updated densities since factoring immediately gives

$$\left|\pi^n_{\text{up}}(\lambda) - \pi_{\text{up}}(\lambda)\right| = \frac{\pi_{\text{init}}(\lambda)}{\pi_{\text{pred}}(Q(\lambda))}\left|\pi^n_{\text{obs}}(Q(\lambda)) - \pi_{\text{obs}}(Q(\lambda))\right|.$$

It then follows that

$$I_{\Lambda\setminus Q^{-1}(A_\delta)} \leq \frac{1}{\delta^{p-1}}\int_{\Lambda\setminus Q^{-1}(A_\delta)} |\pi_{\text{init}}(\lambda)|^p \cdot \frac{\left|\pi^n_{\text{obs}}(Q(\lambda)) - \pi_{\text{obs}}(Q(\lambda))\right|^p}{\pi_{\text{pred}}(Q(\lambda))} d\mu_\Lambda.$$

Utilizing this and a similar argument as before, we obtain

$$I_{\Lambda\setminus Q^{-1}(A_\delta)} \leq \frac{||\pi_{\text{init}}(\lambda)||^{p-1}_{L^\infty(\Lambda)}}{\delta^{p-1}} ||\pi^n_{\text{obs}} - \pi_{\text{obs}}||^p_{L^p(\mathcal{D})}. \tag{E.2}$$

Comparing this to the bound obtained in the previous proof, we note the absence of C^p and that the $L^p(\mathcal{D})$ norm is now of the difference in observed densities as opposed to predicted densities. Both of these differences are attributed to the simpler first step that did not require finding a common denominator. To finish the proof, we simply appeal to the fact that now $\pi^n_{\text{obs}} \to \pi_{\text{obs}}$ in $L^p(\mathcal{D})$ to make the above term small. □

APPENDIX E.4 Proof of Theorem 18: Rate of Convergence with Observed in L^p

Suppose $\pi_{\text{init}} \in L^\infty(\Lambda)$ and π_{obs} are chosen so that Assumption 1 is satisfied. If (π^n_{obs}) satisfies Assumption 3, $\pi^n_{\text{obs}} \to \pi_{\text{obs}}$ in $L^p(\mathcal{D})$, and the convergence rate of $\mathbb{P}^n_{\text{obs}}$ is of order $O(\rho(n))$ on almost all of $\mathcal{D}$, then the convergence rate of $\mathbb{P}^n_{\text{up}}$ is of order $O(\rho(n))$ on almost all of Λ.

Proof. This follows immediately from the bound obtained in Equation (E.2) in the proof of Theorem 17 in Appendix APPENDIX E.3. □

APPENDIX E.5 Code to Reproduce Results

All of the scripts used to generate the numerical results in this paper can be found at

`https://github.com/sandialabs/MrHyDE/tree/main/scripts/DCI/L1-generalization`