SIAM/ASA J. Uncertainty Quantification
Vol. 12, No. 3, pp. 868–900

© 2024 Society for Industrial and Applied Mathematics
and American Statistical Association

Conditional Sampling with Monotone GANs: From Generative Models to Likelihood-Free Inference*

Ricardo Baptista†, Bamdad Hosseini‡, Nikola B. Kovachki§, and Youssef M. Marzouk¶

Abstract. We present a novel framework for conditional sampling of probability measures, using block triangular transport maps. We develop the theoretical foundations of block triangular transport in a Banach space setting, establishing general conditions under which conditional sampling can be achieved and drawing connections between monotone block triangular maps and optimal transport. Based on this theory, we then introduce a computational approach, called monotone generative adversarial networks (M-GANs), to learn suitable block triangular maps. Our algorithm uses only samples from the underlying joint probability measure and is hence likelihood-free. Numerical experiments with M-GAN demonstrate accurate sampling of conditional measures in synthetic examples, Bayesian inverse problems involving ordinary and partial differential equations, and probabilistic image inpainting.

Key words. measure transport, conditional simulation, likelihood-free inference, optimal transport, GANs, normalizing flows

MSC codes. 49Q22, 62G86, 62F15, 60B05

DOI. 10.1137/23M1581546

1. Introduction.

Conditional simulation can be viewed as the process of generating samples from certain "slices" of a probability measure $\nu \in \mathbb{P}(\mathcal{U} \times \mathcal{Y})$. Intuitively, simulating $u \in \mathcal{U}$ conditioned on a given value of $y \in \mathcal{Y}$ amounts to restricting ν along a hyperplane $y = y^*$, renormalizing, and generating samples from the resulting distribution. Conditional sampling problems are ubiquitous in statistics, applied mathematics, and engineering, where u may represent an output or prediction of interest and y may represent a variable that is observed.[1]

Many *supervised learning* algorithms such as ridge, lasso, or neural network regression assume a finite-dimensional parameterization of u and use a statistical model of the observations,

*Received by the editors June 23, 2023; accepted for publication (in revised form) May 29, 2024; published electronically August 9, 2024.
https://doi.org/10.1137/23M1581546
Funding: The first and fourth authors acknowledge support from the US Department of Energy (DOE) AEOLUS Center (award DE-SC0019303) and from the DOE M2dt Center (award DE-SC0023187). The first author also acknowledges support from an NSERC PGS-D fellowship, the Air Force Office of Scientific Research MURI on "Machine Learning and Physics-Based Modeling and Simulation" (award FA9550-20-1-0358), and a Department of Defense (DoD) Vannevar Bush Faculty Fellowship (award N00014-22-1-2790). The second author acknowledges support from the National Science Foundation, research grant NSF-DMS-2208535, "Machine Learning for Bayesian Inverse Problems." The third author acknowledges support from the NVIDIA Corporation through full-time employment.
†California Institute of Technology, Pasadena, CA 91106 USA (rsb@caltech.edu).
‡University of Washington, Seattle, WA 98195 USA (bamdadh@uw.edu).
§NVIDIA, Santa Clara, CA 95051 USA (nkovachki@nvidia.com).
¶Massachusetts Institute of Technology, Cambridge, MA 02139 USA (ymarz@mit.edu).
[1]Our notation here is chosen to be consistent with the literature on Bayesian inverse problems, where y denotes the data and u denotes an unknown of interest.

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$\nu(\cdot|u)$, perhaps paired with some penalization scheme, to construct a point estimator $\hat{u}(y^*)$ of u for any y^*. In the probabilistic setting described above, where the $\mathcal{U}$-marginal $\nu_{\mathcal{U}}$ is naturally interpreted as a prior measure, such point estimators may coincide with the mode of u conditioned on y^* under specific likelihood and prior models [45]. Fully Bayesian methods, however, go further and seek to characterize the entire conditional measure $\nu(\cdot|y^*)$, thereby providing a natural way of quantifying uncertainty in the predicted outputs. Gaussian process regression is a canonical example, where u is an element of an infinite-dimensional space and ν is also Gaussian on the product space.

Inverse problems in the Bayesian setting [55, 104] fall into the aforementioned framework as well; here, one seeks to recover an unknown parameter u from a realization of *indirect* and noisy observations y^*, where u is typically infinite-dimensional. A prototypical inverse problem takes the form

$$\mathcal{L}(u)p = 0, \qquad y = g(p) + \epsilon, \tag{1.1}$$

where $u \in \mathcal{U}$ represents the parameter of interest, $p \in \mathcal{P}$ is a state variable, and $\mathcal{L}(u)$ is an operator acting on p, parameterized by u. Here, $g\colon \mathcal{P} \to \mathcal{Y}$ is an observation operator that extracts $y \in \mathcal{Y}$ from p, and $\epsilon \in \mathcal{Y}$ is a random variable representing observational noise; $\mathcal{U}$, $\mathcal{Y}$, $\mathcal{P}$ are assumed to be Banach spaces. For instance, $\mathcal{L}$ could be a partial differential operator and g could return pointwise evaluations of the PDE solution p, defined with appropriate boundary conditions (see subsection 4.5 for a concrete example). From a probabilistic perspective, (1.1) specifies the conditional distribution $\nu(\cdot|u)$. In the Bayesian setting [104], one also endows u with a prior and thus fully specifies the joint probability measure ν, with the goal of then characterizing the posterior measure $\nu(\cdot|y^*)$.

The common challenge in the applications outlined above is therefore to sample from a conditional measure $\nu(\cdot|y^*)$, as sampling enables the estimation of arbitrary moments or other expectations. Markov chain Monte Carlo (MCMC) algorithms are widely used for this purpose and provide asymptotically exact estimates, but require repeatedly evaluating the likelihood (e.g., solving (1.1) in the case of Bayesian inverse problems); moreover, one must simulate an entirely new Markov chain for each new value of y^*. Also, the performance of most MCMC algorithms is quite sensitive to the choice of prior and likelihood models [26, 43, 46, 47, 107]. These issues often limit the utility of MCMC in large-scale applications. Variational inference (VI) methods [19, 35, 113] offer an alternative to MCMC by approximating the conditional measure $\nu(\cdot|y^*)$ with a measure ν_θ chosen from a certain tractable family parameterized by θ. For example, one can take ν_θ to be the family of Gaussian measures on $\mathcal{U}$ parameterized by their means and covariance operators. While VI can be significantly more efficient than MCMC, the accuracy of VI is very much limited by the quality of the approximating family. (See [19] for a more detailed discussion and for comparisons between MCMC and VI.)

In this article, we present and analyze a novel framework for conditional sampling using transportation of measure. Our methods fall under the umbrella of VI, although the optimization problems and distributional approximations of interest to us are not standard in VI: our family of approximating measures ν_θ comprises the pushforwards of a chosen reference measure by parameterized *block triangular transport maps*; also, we solve optimization problems whose objectives involve statistical divergences inspired by optimal transport (OT) distances, rather than the KL divergence. In this light, our methods are closely related to modern generative

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

models in machine learning (ML), such as generative adversarial networks (GANs) [41] and normalizing flows (NFs) [64]. Another feature differentiating our approach from MCMC and standard VI is the ability to approximate the entire family of conditionals $\nu(\cdot|y)$ by solving a single optimization problem, making it attractive for settings where conditional simulation for a large collection of observation values is desired. A further distinguishing feature is that our approach is entirely data-driven: the approximate conditionals $\nu_\theta(\cdot|y^*)$ are computed only using samples from the joint measure ν.

In the remainder of this section, we give a summary of our main contributions, followed by a review of relevant literature.

1.1. Main contributions. Consider a *reference measure* η and a *target measure* ν, both of which are Borel measures on the separable Banach space $\mathcal{Y} \times \mathcal{U}$. We assume that η is known and can be simulated at low cost; for example, we can choose η to be the standard Gaussian measure whenever $\mathcal{Y}, \mathcal{U}$ are finite-dimensional, or an appropriate Gaussian process in the Banach space setting. Our goal is to generate approximate samples from $\nu(\cdot|y^*)$. To this end, we pose optimization problems of the form

$$\begin{cases} \min\limits_{\mathsf{F},\mathsf{G}} & \mathcal{D}(\mathsf{T}_\sharp\eta, \nu) + \mathcal{R}(\mathsf{T}), \\ \text{s.t.} & \mathsf{T}(y,u) = (\mathsf{F}(y), \mathsf{G}(\mathsf{F}(y),u)), \\ & \mathsf{F}: \mathcal{Y} \to \mathcal{Y}, \quad \mathsf{G}: \mathcal{Y} \times \mathcal{U} \to \mathcal{U}, \end{cases} \tag{1.2}$$

where $\mathcal{D}$ is a statistical divergence on $\mathbb{P}(\mathcal{Y} \times \mathcal{U})$, $\mathcal{R}$ is an appropriate regularization term, and F, G are parametric maps. We make three main contributions in this work:

- We present a theoretical analysis of (1.2) in an idealized setting where $\mathcal{U}$, $\mathcal{Y}$ are Banach spaces and $\mathcal{R} \equiv 0$. Our analysis yields three primary results: (a) If $\mathsf{T}_\sharp\eta = \nu$, then $\mathsf{G}(y^*,\cdot)_\sharp\eta_{\mathcal{U}} = \nu(\cdot|y^*)$, where $\eta_{\mathcal{U}}$ is the $\mathcal{U}$-marginal of η. (b) Under very general conditions on η, ν and for wide choices of $\mathcal{D}$, problem (1.2) has a minimizer $\mathsf{T}^\dagger$ that satisfies $\mathsf{T}^\dagger_\sharp\eta = \nu$. (c) Under appropriate monotonicity constraints on T, and when $\mathcal{Y}$ and $\mathcal{U}$ are finite-dimensional Euclidean spaces, the resulting conditioning map $\mathsf{G}^\dagger(y^*,\cdot)$ is also unique and is the solution to an OT problem (in fact it is a conditional Brenier map [24]). We present these results in section 2.
- Motivated by this theoretical foundation, we present a computational framework called *monotone generative adversarial networks* (M-GANs) that approximates (1.2) in three steps: (a) Take $\mathcal{D}$ to be an approximate Wasserstein-1 type distance; (b) parameterize F, G as neural networks; (c) impose monotonicity on T via the regularization term $\mathcal{R}$. We then solve the resulting optimization problem using stochastic gradient descent to obtain a minimizer $\mathsf{G}^\dagger$. Given a y^*, we then draw samples $u_j \overset{\text{iid}}{\sim} \eta_{\mathcal{U}}$ and evaluate $\mathsf{G}^\dagger(y^*,u_j)$ to obtain approximate samples from $\nu(\cdot|y^*)$. The M-GAN framework is outlined in section 3.
- We evaluate the performance of the M-GAN approach numerically, with experimental settings ranging from low-dimensional synthetic problems to high-dimensional ML applications and infinite-dimensional Bayesian inverse problems involving PDEs. These experiments can be found in section 4.

A core feature of the M-GAN framework is that to solve (1.2) numerically, we only require samples from the joint measure ν, yet the map $\mathsf{G}^\dagger$ characterizes all of the conditionals

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$\nu(\cdot|y)$. In other words, M-GAN is entirely data-driven and does not require evaluations of a likelihood function or prior density; more generally, it does not require any explicit knowledge or modeling assumptions on the relationship between u and y. Moreover, since the computed $\mathsf{G}^\dagger$ can be evaluated at multiple values of y^* without any additional optimization, the cost of inference is "amortized" over y^* [38, 90, 113, 27, 70].

1.2. Relevant literature. Conditional sampling is an active area of research in computational statistics, ML, and applied mathematics. Conventional methods such as MCMC and VI, as mentioned earlier, have a rich and active literature but a thorough review of these topics is outside the present scope. Instead we focus on literature pertaining to conditional sampling and transportation of measure.

1.2.1. Measure transport in uncertainty quantification. The use of transport maps for conditional sampling has been explored in the uncertainty quantification and inverse problems communities [74, 33, 102, 98]. For problems in Bayesian inference and ML [23, 53, 85], a common approach is to seek monotone triangular maps that approximate the classic Knothe–Rosenblatt (KR) rearrangement [95]. By construction, components of the KR rearrangement push forward a product reference measure η to the target conditionals, which is precisely what is desired for conditional sampling. While the KR rearrangement can be written explicitly in terms of marginal-conditional distribution and quantile functions, direct computation using this definition is typically infeasible. coordinate bases. We demonstrateThe approach of [74, 33, 102] instead is to formulate problems akin to (1.2) by choosing $\mathcal{D}$ to be the KL divergence, taking the reference η to be the standard Gaussian, and parameterizing T in a space of monotone triangular functions. These choices naturally affect the accuracy of the resulting transport. Also, most triangular map representations (with the exception of [111]) are limited to finite-dimensional input spaces $\mathcal{U}$ and $\mathcal{Y}$, and the fully triangular form of T requires selecting a particular ordering of the coordinate bases. We demonstrate in subsection 4.2 that this choice can have a significant impact on accuracy in practice. Furthermore, monotone parameterizations of T can lead to poorly behaved optimization problems (e.g., with many local minima) unless one exercises sufficient care, as described in [13]. Our M-GAN framework addresses the aforementioned drawbacks of triangular transport maps by generalizing the formulation of [74, 33, 102] in several ways: (a) we allow wider choices of $\mathcal{D}$; (b) we ask only for T to be block triangular, such that no ordering of coordinate bases for $\mathcal{U}$ or $\mathcal{Y}$ is needed; and (c) we establish validity of our formulation on infinite-dimensional Banach spaces. To achieve conditional sampling, we do not even require T to be monotone, although we do impose monotonicity in practice and enforce it in some of our theoretical results (e.g., in making a link to OT).

Analysis of triangular transport maps is a classical topic going back to the works of Knothe [63] and Rosenblatt [93]. Basic properties of such maps, such as existence, uniqueness, and regularity, have since been studied in general settings including infinite-dimensional Banach spaces [22, 21]. Applied analysis of triangular maps, pertaining to algorithms, has become of interest much more recently: [110, 111] show that under appropriate assumptions on the reference and target measures, the KR rearrangement is analytic on the finite- or infinite-dimensional hypercube and can be well approximated with sparse polynomials or deep ReLU networks; [50] considers a variational characterization of the KR rearrangement akin to (1.2)

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

and studies the statistical consistency and convergence of the empirical approximation of the map given samples from the target ν; [108] establishes optimal minimax rates of convergence for nonparametric density estimators based on triangular and other transport maps, by adapting techniques from M-estimation and empirical process theory to the transport setting; [52] analyzes the tail behavior of triangular maps, revealing an intricate balance between the tails of the reference and target measures and the expressive power of Lipschitz maps; and [29] presents an efficient and scalable tensor train parameterization of triangular maps for conditional sampling.

Our theoretical contributions are distinct from the aforementioned articles in four aspects: (a) we do not limit ourselves to triangular/KR maps and instead consider the much more general problem in (1.2); (b) we allow for a generic choice of $\mathcal{D}$ as opposed to the KL divergence; (c) we develop existence and convergence results for the conditioning map G as opposed to the full map T; and (d) we connect our block triangular construction to recent results in OT.

1.2.2. Measure transport in ML. Measure transport problems have also attracted considerable interest in the ML literature, particularly for generative modeling [80, 54]. Following [54], we say that an ML model or algorithm is "generative" if it characterizes the joint measure ν rather than the conditional measure $\nu(\cdot|y^*)$. In this definition, the map T obtained by solving (1.2) is a generative model. Popular generative models of relevance to our M-GAN framework are GANs [41, 40], NFs [90, 84, 64], and, to some extent, variational autoencoders (VAEs) [62, 32]. All of these methods and their variants solve problems of a form similar to (1.2), but with three core differences: (a) the reference measure η in GANs and VAEs is often defined on a lower-dimensional space, enabling natural dimension reduction; (b) the map T is parameterized directly and the maps F, G are omitted from the formulation; and (c) the regularization term R is not identified, or its impact is not analyzed explicitly. In GANs, the map T is often parameterized by a single neural network and $\mathcal{D}$ is taken to be a GAN loss function, which can be viewed as an approximation to a Wasserstein-type distance or a variational form of a statistical divergence [81, 10]. NFs represent T as a *composition* of invertible and often *triangular* [53] neural networks, typically interleaved with permutations, and may choose $\mathcal{D}$ to be the KL divergence [84]. In this light, NFs are closely related to the triangular maps of [74, 33, 102, 98].[2] VAEs pose a slightly different problem to GANs and NFs by parameterizing both T and its inverse T^{-1} as separate neural networks and approximating them simultaneously. The KL divergence is again employed in most VAE applications.

As the name M-GAN suggests, our proposed framework is closely related to GANs. In fact, one can view M-GAN as a combination of GANs and NFs with a particular parameterization of the map T. However, we emphasize that the aforementioned generative models aim to approximate the map T, with the ultimate goal of sampling the joint distribution ν. The task of conditioning ν is not of direct interest and is often tackled in a secondary step using Bayes' rule or other standard (or ad hoc) conditioning techniques. Thus, a defining feature of M-GAN is that it allows us to directly characterize the family of conditionals $\nu(\cdot|y)$ through the map G. We then obtain the generative model T as a by-product.

[2] In fact, this connection to triangular maps and our analysis in section 2 imply that NFs can easily be retrofitted for conditional sampling, simply by appending the conditioning variables y as additional inputs and constraining the permutation layers appropriately [27, 88, 9, 74].

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

Several previous efforts to adapt generative models for conditional sampling exist in the ML literature. Most notably, [76, 51] define conditional GANs and VAEs by training neural networks that depend on the joint variables (y, u) to obtain maps that can generate samples from multimodal distributions. However, these formulations are limited to settings where y is a discrete variable and many u samples are available for a given y; such data is typically not available for continuous y.

The articles [51, 15] addressed the more difficult problem of approximating all the conditionals of the joint distribution ν by employing a weighted loss function over all possible choices of conditionals. The article [103] also considers the related problem of estimating arbitrary conditional densities using an energy-based model. These approaches have two major drawbacks: the loss does not guarantee that any particular conditional is obtained correctly/accurately, and the problem quickly becomes infeasible in high- or infinite-dimensional settings. [109] considered the problem of correctly extracting a single conditional from a fixed generative model using a VI loss. The proposed method must be retrained for *each* new value of y, in contrast to M-GAN where the map $\mathsf{G}^\dagger$ characterizes the entire conditional family $\nu(\cdot|y)$ simultaneously.

The articles [1, 89, 114] are closest to our construction. The approaches of [1, 89] are similar to each other and can be viewed as particular versions of M-GAN by taking F to be the identity map, omitting the monotonicity penalty/constraint on T, and choosing a particular form of $\mathcal{D}$ and G. The article [114] considers a similar situation by assuming η to be Gaussian, choosing $\mathcal{D}$ to be an f-divergence, and parameterizing G with a neural network. The theoretical exposition in [114] is well aligned with our results in section 2, although they only consider the case where $\mathcal{Y}$,$\mathcal{U}$ are finite-dimensional Euclidean spaces and do not make the connection to OT. Our work can be differentiated from these efforts in three directions: (a) our theoretical results and algorithms are valid on infinite-dimensional Banach spaces, a setting that is crucial for PDE inverse problems; (b) our formulation is more general, placing minimal assumptions on $\mathcal{D}$, the parameterization of T, or the choice of reference distribution η; (c) by including a monotonicity penalty, we are able to provide further understanding of the solution of (1.2) by connecting our minimizer to OT.

We briefly mention other relevant works at the intersection of conditional simulation, generative modeling, and statistical inference. The works [18, 94, 8, 97, 5] use parametric or nonparametric models for density estimation although they mainly focus on structural constraints on target conditional densities and do not focus on conditional *sampling* as we do here. The article [44] considers a GAN for conditional simulation in the setting where a priori information about the moments of the conditional measure is available and utilizes this information to improve the quality of the generative model. The article [77] introduces a specialized discrepancy that measures the quality of a conditional GAN after training, while [36, 100, 86] leverage generative models to enhance the convergence properties of MCMC algorithms. Finally, we note that probabilistic diffusion models have recently been adapted for the solution of inverse problems [101, 14, 99] which take a different approach to transport maps by learning a (possibly problem-agnostic) diffusion model given prior samples and appropriately guiding the reverse process to generate samples from the conditional of interest.

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

1.2.3. Connection to OT. Given two measures $\eta, \nu \in \mathbb{P}(\mathcal{Z})$, the Monge problem of optimal transportation seeks maps T that satisfy $\mathsf{T}_\sharp \eta = \nu$ while minimizing functionals of the form $\int c(z, \mathsf{T}(z))\eta(\mathrm{d}z)$ for appropriate cost functions $c : \mathcal{Z} \times \mathcal{Z} \to \mathcal{R}$. Our analysis and the construction of the M-GAN framework are strongly inspired by OT. In fact, existence and uniqueness results from OT can be extended to minimizers of (1.2). Furthermore, the monotonicity penalty in the M-GAN framework is directly motivated by uniqueness results for the well-known Brenier maps [75] and their block triangular extensions [24]. The articles [25, 79] are also closely related to our (block) triangular constructions. We present a more detailed discussion of how our approach relates to OT in subsection 2.3.

We note, however, that there are fundamental differences between problem (1.2) and the OT problem. Most importantly, OT maps are constrained to push the reference η to the target ν exactly while minimizing a transport cost; instead, we ask only to minimize $\mathcal{D}(\mathsf{T}_\sharp \eta, \nu)$ and thus may not match the target measure ν exactly. Furthermore, we restrict the function space to which T belongs and regularize this map via the penalty R. These relaxations allow us to obtain "nicer" transport maps that can be computed in a stable manner. Despite these relaxations, we demonstrate in subsection 4.3 that M-GAN maps converge to certain OT maps if $\mathcal{D}$ and the space in which one seeks T are chosen correctly. Our results in this direction draw on results from [24]. This observation suggests that M-GAN can serve as a numerical method for approximating (conditional) OT maps, which is a topic that has attracted much interest in the ML community [37, 96, 66]. We also mention recent articles [59, 105, 3] that use conditional OT strategies resembling M-GANs for filtering and data assimilation.

1.3. Outline. The remainder of this paper is organized as follows. Section 2 establishes the necessary conditions for performing conditional sampling via block triangular transport maps, and discusses the existence and uniqueness of these maps in relation to those found via OT. Section 3 presents our framework for monotone transport map approximation. Section 4 presents numerical results for generative modeling and the solution of Bayesian inverse problems, followed by a concluding discussion in section 5.

2. Theoretical foundations.

In this section, we develop a theoretical analysis of problem (1.2) in idealized settings, which serves as a foundation for the M-GAN framework introduced in section 3.

Let $\mathcal{U}$, $\mathcal{V}$, $\mathcal{Y}$, $\mathcal{W}$ be separable Banach spaces with Borel σ-algebras $\mathcal{B}(\mathcal{U})$, $\mathcal{B}(\mathcal{V})$, $\mathcal{B}(\mathcal{Y})$, $\mathcal{B}(\mathcal{W})$, respectively. Define the product spaces $\mathcal{Z} := \mathcal{Y} \times \mathcal{U}$ and $\mathcal{S} := \mathcal{W} \times \mathcal{V}$, with corresponding product Borel σ-algebras $\mathcal{B}(\mathcal{Z})$ and $\mathcal{B}(\mathcal{S})$. Let $\mathbb{P}(\mathcal{U})$, $\mathbb{P}(\mathcal{Y})$, $\mathbb{P}(\mathcal{Z})$, $\mathbb{P}(\mathcal{V})$, $\mathbb{P}(\mathcal{W})$, $\mathbb{P}(\mathcal{S})$ denote spaces of Borel probability measures on their respective Banach spaces. For a measure $\mu \in \mathbb{P}(\mathcal{Z})$ (resp., $\in \mathbb{P}(\mathcal{S})$) we use $\mu_{\mathcal{U}}$ and $\mu_{\mathcal{Y}}$ (resp., $\mu_{\mathcal{V}}$ and $\mu_{\mathcal{W}}$) to denote the marginals of μ on $\mathcal{U}$ and $\mathcal{Y}$ (resp., $\mathcal{V}$ and $\mathcal{W}$). Finally, for any set $B \in \mathcal{Z}$ we define the slice $B_y := \{u : (u, y) \in B\}$. In what follows, $\mathcal{U}$ will represent the parameter space and $\mathcal{Y}$ will represent the space of data on which we condition, with $\mathcal{V}$ and $\mathcal{W}$ being their corresponding "reference" spaces. We now recall the definition of regular conditional measures:

Definition 2.1 (regular conditional measures). *Let $\mu \in \mathbb{P}(\mathcal{Z})$. We say $\mu(\cdot|y)$ is a system of regular conditional measures for μ if*

1. *$\forall y \in \mathcal{Y}$, $\mu(\cdot|y)$ is a probability measure on $\mathcal{B}(\mathcal{U})$;*

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

2. $\forall A \in \mathcal{B}(\mathcal{U})$, *the function* $y \mapsto \mu(A|y)$ *is measurable with respect to* $\mathcal{B}(\mathcal{Y})$ *and is* $\mu_{\mathcal{Y}}$*-integrable;*
3. $\forall B \in \mathcal{B}(\mathcal{Z})$, *it holds that* $\mu(B) = \int_{\mathcal{Y}} \mu(B_y|y)\mu_{\mathcal{Y}}(\mathrm{d}y)$.

In what follows, we often refer to systems of regular conditional measures simply as systems of conditional measures or "conditionals." By [20, Cor. 10.4.15] we have the following existence and uniqueness result.

Proposition 2.2. *Consider the above setting with* $\mu \in \mathbb{P}(\mathcal{Z})$. *Then the following hold:*
(a) *(Existence) There exist Radon conditional measures* $\mu(\cdot|y)$ *of* μ $\forall$ $y \in \mathcal{Y}$.
(b) *(Uniqueness) The conditional measures* $\mu(\cdot|y)$ *are unique up to* $\mu_{\mathcal{Y}}$ *null sets.*

Now consider a *reference measure* $\eta = \eta_{\mathcal{W}} \otimes \eta_{\mathcal{V}} \in \mathbb{P}(\mathcal{S})$, that is, of product form, and a *target measure* $\nu \in \mathbb{P}(\mathcal{Z})$. Our goal throughout this article is to characterize the conditionals $\nu(\cdot|y)$ via a transformation of the reference measure η—specifically, a transformation of the marginal $\eta_{\mathcal{W}}$. To this end, consider a *block triangular map* of the form

$$\mathsf{T} : \mathcal{S} \to \mathcal{Z}, \qquad \mathsf{T}(w, v) = (\mathsf{F}(w), \mathsf{G}(\mathsf{F}(w), v)), \tag{2.1}$$

which in turn is defined through the maps

$$\mathsf{F} : \mathcal{W} \to \mathcal{Y}, \qquad \mathsf{G} : \mathcal{Y} \times \mathcal{V} \to \mathcal{U}. \tag{2.2}$$

Remark 2.3. We refer to T as a *block triangular map* since, in the setting where $\mathcal{V}$, $\mathcal{U}$, $\mathcal{W}$, and $\mathcal{Y}$ are finite-dimensional Euclidean spaces, the Jacobian matrix of T is block triangular. We note that such maps are also simply called triangular in the literature; see, for example, [20, sect. 10.10(vii)]. However, we prefer the term block triangular to set our parameterizations apart from strictly triangular maps such as the KR rearrangement considered in [74] or the elementary maps in NFs [64, 84]. We demonstrate in subsection 4.2 that block triangular maps perform quite differently from strictly triangular maps in practice.

2.1. Block triangular transport.

The following theorem is the cornerstone of our methodology for approximating the conditionals of ν via block triangular transport.

Theorem 2.4. *Consider a reference* $\eta = \eta_{\mathcal{W}} \otimes \eta_{\mathcal{V}} \in \mathbb{P}(\mathcal{S})$ *and a target* $\nu \in \mathbb{P}(\mathcal{Z})$ *and let* T *be a block triangular map of the form* (2.1) *satisfying* $\mathsf{T}_\sharp \eta = \nu$. *Then for* $\mathsf{F}_\sharp \eta_{\mathcal{W}}$*-a.e.* y *it holds that* $\mathsf{G}(y, \cdot)_\sharp \eta_{\mathcal{V}} = \nu(\cdot|y)$.

Proof. Consider the maps $\widetilde{\mathsf{T}} \colon (y, v) \mapsto (y, \mathsf{G}(y, v))$ and $(\mathsf{F} \times \mathrm{Id}) \colon (w, v) \mapsto (\mathsf{F}(w), v)$ and observe that $\mathsf{T} = \widetilde{\mathsf{T}} \circ (\mathsf{F} \times \mathrm{Id})$. Let $B \in \mathcal{B}(\mathcal{Z})$. We have by the hypothesis of the theorem that

$$\begin{aligned} \int_B \nu(\mathrm{d}z) &= \int_B \mathsf{T}_\sharp \eta(\mathrm{d}z) = \int_{\widetilde{\mathsf{T}}^{-1}(B)} (\mathsf{F} \times \mathrm{Id})_\sharp (\eta_{\mathcal{W}} \otimes \eta_{\mathcal{V}})(\mathrm{d}w, \mathrm{d}v) \\ &= \int_{\widetilde{\mathsf{T}}^{-1}(B)} (\mathsf{F}_\sharp \eta_{\mathcal{W}} \otimes \eta_{\mathcal{V}})(\mathrm{d}y, \mathrm{d}v) = \int_{\widetilde{\mathsf{T}}^{-1}(B)} (\nu_{\mathcal{Y}} \otimes \eta_{\mathcal{V}})(\mathrm{d}y, \mathrm{d}v), \end{aligned}$$

where the last identity follows from the observation that $\mathsf{T}_\sharp \eta = \nu$ implies $\mathsf{F}_\sharp \eta_{\mathcal{W}} = \nu_{\mathcal{Y}}$ due to the product structure of η. Now observe that $\widetilde{\mathsf{T}}^{-1}(B)_y = \{u : (y, u) \in \widetilde{\mathsf{T}}^{-1}(B)\} = \mathsf{G}(y, \cdot)^{-1}(B_y)$.

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

We can continue the above calculation as follows:

$$\int_B \nu(\mathrm{d}z) = \int_{\mathcal{Y}} \eta_{\mathcal{V}}(\widetilde{\mathsf{T}}^{-1}(B)_y)\nu_{\mathcal{Y}}(\mathrm{d}y) = \int_{\mathcal{Y}} \eta_{\mathcal{V}}(\mathsf{G}(y,\cdot)^{-1}(B_y))\nu_{\mathcal{Y}}(\mathrm{d}y) = \int_{\mathcal{Y}} \mathsf{G}(y,\cdot)_\sharp \eta_{\mathcal{V}}(B_y)\nu_{\mathcal{Y}}(\mathrm{d}y).$$

Thus we have established that $\mathsf{G}(y,\cdot)_\sharp\eta_{\mathcal{V}}$ is a conditional measure for ν. The desired result now follows from Proposition 2.2(b) stating the essential uniqueness of conditional measures. ∎

We highlight the simplicity of the proof and the generality of the conditions in Theorem 2.4. We do not require any assumptions such as regularity or monotonicity of the map T beyond the parameterization (2.1), and we certainly do not require η and ν to be defined on the same spaces. The main restriction of our result is the assumption that $\eta = \eta_{\mathcal{W}} \otimes \eta_{\mathcal{V}}$, but this is an assumption on the reference measure, which can be chosen with considerable freedom.

The existence of the maps F and G may appear nontrivial at first sight. However, upon inspection of the proof of Theorem 2.4 we realize that the map F is to some extent innocuous. For example, we can simply choose $\mathcal{W} = \mathcal{Y}$ and $\eta = \nu_{\mathcal{Y}} \otimes \eta_{\mathcal{V}}$ and take F to be the identity map. Indeed, this is the approach taken in [24, 89, 114]. If the above choice for η is infeasible we can still take F to be any map that transports $\eta_{\mathcal{W}}$ to $\nu_{\mathcal{Y}}$. An obvious choice would be a Brenier OT map, which exists under very general assumptions on $\eta_{\mathcal{W}}$ [106]. Thus we focus our attention on the map G for the remainder of this subsection. First, we demonstrate that G can be identified in closed form in certain settings.

Example 2.5 (Gaussian random variables). Let η and ν be multivariate Gaussian measures where the conditional distribution $\nu(\cdot|y)$ has mean m_y and covariance Σ_y and $\eta_{\mathcal{V}}$ is standard Gaussian. Let G be the affine transport map $\mathsf{G}(y,v) = m_y + \Sigma_y^{1/2} v$, where $\Sigma_y^{1/2}$ is any matrix square root. Then, $v \mapsto \mathsf{G}(y,v)$ pushes forward $\eta_{\mathcal{V}}$ to the conditional measure $\nu(\cdot|y)$.

Example 2.6 (invertible transformations of random variables). Consider measures of the form $\nu = \mathrm{Law}\{(y,u)|y = h(u-\xi),\ u \sim \pi_1,\ \xi \sim \pi_2\}$, where $\pi_1, \pi_2 \in \mathbb{P}(\mathcal{U})$. If h^{-1} exists, then we can readily verify that $u = h^{-1}(y) + \xi$. In other words, $\nu(u|y) = \pi_2(u - h^{-1}(y))$. Now take $\eta = \nu_{\mathcal{Y}} \otimes \pi_2$ and observe that $\mathsf{G}(y,v) = v + h^{-1}(y)$ is the desired transport map.

A natural question arises regarding the existence of G. While identification of G is nontrivial, the existence of such a map can be guaranteed under very general assumptions, following classic results from probability theory. Below we consider the case where $\mathsf{F} = \mathrm{Id}$.

Proposition 2.7. *Take $\eta = \nu_{\mathcal{Y}} \otimes U[0,1]$ where $U[0,1]$ denotes the uniform measure on the interval $[0,1]$. Then there exists a measurable map $\mathsf{G} : \mathcal{Y} \times [0,1] \to \mathcal{U}$ so that $\mathsf{G}(y,\cdot)_\sharp U[0,1] = \nu(\cdot|y)$ for any target $\nu \in \mathbb{P}(\mathcal{Z})$.*

Proof. Recall that we equipped $\mathcal{Z}$ with the Borel σ-algebra and that the system of conditional measures $\nu(\cdot|y)$ are by definition transition kernels. Then a direct application of [58, Lem. 2.22] gives the desired result. ∎

We can extend the above result by replacing the uniform measure $U[0,1]$ with another Radon measure $\eta_{\mathcal{V}} \in \mathbb{P}(\mathcal{V})$.[3] This uses the fact that Borel measures on Polish spaces are

[3]Indeed [114] presents a similar extension for the case of a Gaussian reference measure following the same line of thinking.

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

isomorphic to Borel measures on $[0,1]$; see [58, Chap. 1 and Thm. A1.6]. Thus, the existence of G is not an issue in our separable Banach space setting. More delicate questions arise, however, such as the characterization and regularity of the maps G (and subsequently T), which are important for approximation. We address some of these questions in the next subsection.

2.2. Variational characterization of block triangular maps. We now turn our attention to identifying block triangular maps T of the form (2.1) via variational formulations, with a view toward practical algorithms. Any computational approach will require approximation of the map T with a sequence of maps T^n and so we need a notion of convergence for the resulting pushforwards to the conditional measures $\nu(\cdot|y)$. We obtain such a result below under the assumption that the target ν is nondegenerate. Recall the following definition.

Definition 2.8. *A measure $\mu \in \mathbb{P}(\mathcal{Z})$ is nondegenerate if for any collection of bounded linear functionals $\ell_1, \dots, \ell_n \in \mathcal{Z}^*$ (the dual of $\mathcal{Z}$) the measures $(\ell_1, \dots, \ell_n)_\sharp \mu \in \mathbb{P}(\mathbb{R}^n)$ are absolutely continuous.*

We now obtain the following convergence result in the setting where the approximating sequence $\mathsf{T}^n_\sharp \eta$ converges to ν weakly on the product space.

Theorem 2.9. *Consider a reference measure $\eta = \eta_{\mathcal{W}} \otimes \eta_{\mathcal{Y}} \in \mathbb{P}(\mathcal{S})$ and a nondegenerate target measure $\nu \in \mathbb{P}(\mathcal{Z})$. Let $\{\mathsf{T}^n\}_{n\geq 0}$ be a sequence of maps of the form (2.1) with component maps $\mathsf{F}^n, \mathsf{G}^n$ as in (2.2). Furthermore, suppose that $\mathsf{T}^n_\sharp \eta \to \nu$ weakly as $n \to \infty$. Then, for any $r > 0$, $y^* \in \mathcal{Y}$, and $f \in C_b(\mathcal{U})$ it holds that*

$$\lim_{n\to\infty} \int_{B_r(y^*)} \int_{\mathcal{U}} f(u) \mathsf{G}^n(y,\cdot)_\sharp \eta_{\mathcal{W}}(\mathrm{d}u) \mathsf{F}^n_\sharp \eta_{\mathcal{W}}(\mathrm{d}y) \to \int_{B_r(y^*)} \int_{\mathcal{U}} f(u) \nu(\mathrm{d}y, \mathrm{d}u),$$

where $C_b(\mathcal{U})$ denotes the space of continuous and bounded functions on $\mathcal{U}$.

Proof. By definition of weak convergence we have for $(g, f) \in C_b(\mathcal{Y}) \times C_b(\mathcal{X})$ that

$$\int_{\mathcal{Y}} g(y) \int_{\mathcal{U}} f(u) \mathsf{G}^n(y,\cdot)_\sharp \eta_{\mathcal{V}}(\mathrm{d}u) \mathsf{F}^n_\sharp \eta_{\mathcal{W}}(\mathrm{d}y) \to \int_{\mathcal{Y}} g(y) \int_{\mathcal{U}} f(u) \nu(\mathrm{d}y, \mathrm{d}u).$$

It further follows from [2, Lem. 6.1] that since ν is nondegenerate, then open and convex sets are continuity sets of ν. The desired result then follows from an application of the Portmanteau theorem [20, Cor. 8.2.10]. ∎

We view the above theorem as an averaged weak convergence for the conditionals, i.e., integrals of bounded and continuous functions with respect to $\mathsf{G}^n(y^*,\cdot)_\sharp \eta_{\mathcal{V}}$ converge to integrals with respect to $\nu(\cdot|y^*)$ so long as we average those integrals over small balls around y^*. One can obtain stronger convergence results such as y-a.s. convergence of $\mathsf{G}^n(y,\cdot)_\sharp \eta_{\mathcal{V}}$ to $\nu(\cdot|y)$ by imposing stronger conditions on ν using general convergence results for conditional measures; see [28, 39]. We do not pursue this direction at the moment since the required conditions on ν are difficult to verify in practice.

We also note that the assumption of nondegeneracy on the target ν can be replaced with other (possibly more relaxed) conditions. For example, we only need $B_r(y^*)$ to be a continuity

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

set of $\nu_{\mathcal{Y}}$. Alternatively, if ν is degenerate we can always relax the statement of Theorem 2.9 to a convergence result of the form

$$\lim_{n\to\infty}\int_{\mathcal{Y}} g(y)\int_{\mathcal{U}} f(u)\mathsf{G}^n(y,\cdot)_\sharp\eta_{\mathcal{V}}(\mathrm{d}u)\mathsf{F}^n_\sharp\eta_{\mathcal{W}}(\mathrm{d}y)\to\int_{\mathcal{Y}} g(y)\int_{\mathcal{U}} f(u)\nu(\mathrm{d}y,\mathrm{d}u),$$

where g is a Lipschitz approximation to the indicator of $B_r(y^*)$—for example,

$$g(y):=\begin{cases}1, & y\in B_r(y^*),\\ \max\left\{0,1-\dfrac{\|y-y^*\|_{\mathcal{Y}}-r}{\epsilon}\right\}, & y\in B_r(y^*)^c,\end{cases}$$

for a small parameter $\epsilon>0$.

The above approximation result motivates a variational characterization of the block triangular maps T (along with their approximations T^n) by minimizing statistical divergences. If the chosen divergence metrizes weak convergence, one can then directly apply Theorem 2.4 to obtain convergence of expected values of quantities of interest. This line of thinking leads us to optimization problems of the form (1.2). We recall the definition of a statistical divergence.

Definition 2.10. *A function $\mathcal{D}\colon\mathbb{P}(\mathcal{Z})\times\mathbb{P}(\mathcal{Z})\to\mathbb{R}$ is called a statistical divergence (or simply a divergence) on $\mathbb{P}(\mathcal{Z})$ if for $\mu_1,\mu_2\in\mathbb{P}(\mathcal{Z})$ it holds that*

1. $\mathcal{D}(\mu_1,\mu_2)\geq 0$,
2. $\mathcal{D}(\mu_1,\mu_2)=0$ *if and only if* $\mu_1=\mu_2$.

We now pose the following optimization problem:

$$\underset{\mathsf{T}\in\mathcal{T}}{\text{minimize}}\,\mathcal{D}(\mathsf{T}_\sharp\eta,\nu), \tag{2.3}$$

where $\mathcal{T}$ is the space of measurable maps T parameterized as in (2.1) and (2.2). That is,

$$\mathcal{T}:=\{\mathsf{T}:\mathcal{S}\to\mathcal{Z}\,:\,\mathsf{T}(w,v)=(\mathsf{F}(w),\mathsf{G}(\mathsf{F}(w),v))\text{ for }\mathsf{F}:\mathcal{W}\to\mathcal{Y},\ \mathsf{G}:\mathcal{Y}\times\mathcal{V}\to\mathcal{U}\}. \tag{2.4}$$

Remark 2.11. It follows from Proposition 2.7 and the subsequent discussion that problem (2.3) has a global minimizer $\mathsf{T}^\dagger$ achieving $D(\mathsf{T}^\dagger_\sharp\eta,\nu)=0$ so long as we take $\eta=\eta_{\mathcal{W}}\otimes\eta_{\mathcal{V}}$. These minimizers, however, are not unique. For example, if we take $\mathcal{S}=\mathcal{Z}$ to be finite-dimensional Euclidean spaces with an atomless reference measure η, then the KR rearrangement T_{KR} serves as a global minimizer of (2.3). At the same time, letting P be a block-diagonal permutation matrix that reorders the $\mathcal{U}$ (similarly the $\mathcal{Y}$) coordinates of $\mathcal{Z}$, we can also construct the KR rearrangement $\mathsf{T}'_{\mathrm{KR}}$ between the measures $P_\sharp\eta$ and $P_\sharp\nu$, and then $P^{-1}\circ\mathsf{T}'\circ P$ will also be a minimizer of (2.3). Another example is the conditional Brenier maps of Proposition 2.12, which are fully block triangular as opposed to the KR maps that are strictly triangular.

2.3. Monotone block triangular maps. We now consider restricting the set $\mathcal{T}$ in problem (2.3) in a way that leads to unique minimizers, which also have the desirable regularity properties of OT maps. To this end, we restrict our attention to the setting where $\mathcal{W}=\mathcal{Y}$, $\mathcal{V}=\mathcal{U}$, and hence $\mathcal{S}=\mathcal{Z}=\mathcal{Y}\times\mathcal{U}$, i.e., the reference measure η and the target measure ν are defined on the same space. We also let $\mathcal{Y}$ and $\mathcal{U}$ be finite-dimensional Euclidean spaces. Motivated by [24] and existing literature on approximations of the KR rearrangement [74], we consider two subsets of $\mathcal{T}$:

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

- the set $\mathcal{T}^M \subset \mathcal{T}$ of monotone maps,

$$\mathcal{T}^M := \left\{ \mathsf{T} \in \mathcal{T} : \left(\mathsf{T}(z) - \mathsf{T}(z')\right)^\top (z - z') \geq 0 \ \ \forall z, z' \in \mathcal{Z} \right\};$$

- the set $\mathcal{T}^B \subset \mathcal{T}$ of maps for which F and G are gradients of convex functions,

$$\begin{aligned} \mathcal{T}^B := \{ \mathsf{T} \in \mathcal{T} : & \exists f : \mathcal{Y} \to \mathcal{R}, \ \ \exists g : \mathcal{Y} \times \mathcal{U} \to \mathcal{R} \text{ such that} \\ & y \mapsto f(y) \text{ and } v \mapsto g(y, v) \text{ are convex and} \\ & \mathsf{F}(y) = \nabla_y f(y), \ \ \mathsf{G}(y, v) = \nabla_v g(y, v) \}. \end{aligned}$$

The spaces $\mathcal{T}^M$ and $\mathcal{T}^B$ are closely related but are not the same. Elements of $\mathcal{T}^B$ are monotone, but $\mathcal{T}^B \subset \mathcal{T}^M$ due to a well-known result of Rockafellar stating that maximal cyclically monotone maps T are uniquely determined by gradients of proper convex functions [91, Thm. 24.8, 24.9]. Cyclic monotonicity is a stronger condition than monotonicity and so there are monotone maps that are not gradients of convex functions.

In subsection 3.2 we develop regularization techniques using the space $\mathcal{T}^M$, but we note that $\mathcal{T}^B$ is more convenient for our theoretical analysis. Using $\mathcal{T}^M$ leads to a natural penalty term that is convenient in practice and can be implemented with minimal restrictions on our parameterization of the maps, while $\mathcal{T}^B$ motivates the use of parameterizations for convex functions [7]. In either case, the monotonicity of the minimizers leads to desirable uniqueness and regularity properties. In particular, the Browder–Minty theorem tells us that continuous, bounded, and coercive monotone maps are surjective [112], and this surjectivity allows us to overcome issues with overfitting, which is also referred to as *mode collapse* in the generative modeling literature. We now show a uniqueness result for (2.3) when the space $\mathcal{T}^B$ is utilized. Let us start by recalling [24, Thm. 2.3]. We emphasize that we only consider the case where $\mathcal{V} = \mathcal{U}$, and hence eliminate the notation for the reference space of the parameter.

Proposition 2.12. *Let $\mathcal{Y}$ and $\mathcal{U}$ be finite-dimensional Euclidean spaces. Consider a target measure $\nu \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$ and a reference measure $\eta \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$. Assume the following:*

(i) *The reference measure $\eta \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$ has the form $\eta = \nu_{\mathcal{Y}} \otimes \eta_{\mathcal{U}}$.*
(ii) *The reference marginal $\eta_{\mathcal{U}}$ has a Lebesgue density with convex support on $\mathcal{U}$.*
(iii) *For each $y \in \mathcal{Y}$, the target conditional measure $\nu(\cdot|y)$ admits a Lebesgue density.*
(iv) *$\int_{\mathcal{Y}} \int_{\mathcal{U}} \|u\|^2 \nu(\mathrm{d}y, \mathrm{d}u) < \infty$ and $\int_{\mathcal{U}} \|u\|^2 \eta_{\mathcal{U}}(\mathrm{d}u) < \infty$.*

Then there exists a unique *map $\mathsf{G} = \nabla_u g(y, u)$, where $u \mapsto g(y, u)$ is convex $\forall y \in \mathcal{Y}$, such that*

$$\mathsf{G}(y, \cdot)_\sharp \eta_{\mathcal{U}} = \nu(\cdot|y) \quad \text{for } \nu_{\mathcal{Y}}\text{-a.e. } y. \tag{2.5}$$

Moreover, this G minimizes the quadratic cost $\mathcal{M}(\mathsf{S}) := \mathbb{E}_{(y,u)\sim\eta} \|u - \mathsf{S}(y, u)\|^2$ among all maps $\mathsf{S} : \mathcal{Y} \times \mathcal{U} \to \mathcal{U}$ that satisfy (2.5).

Following [24], we call the map identified by Proposition 2.12 a "conditional Brenier map." Next, combining this result with Theorem 2.4, we obtain a uniqueness result for optimization problems of the form (2.3) over the set $\mathcal{T}^B$.

Theorem 2.13. *Let $\mathcal{Y}$ and $\mathcal{U}$ be finite-dimensional Euclidean spaces. Consider a target measure $\nu \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$ and a reference measure $\eta \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$ of the form $\eta = \eta_{\mathcal{Y}} \otimes \eta_{\mathcal{U}}$. Suppose*

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

$\eta_{\mathcal{Y}}$ has no atoms and that conditions (ii)–(iv) *of Proposition* 2.12 *are satisfied. Then, the optimization problem*

$$\underset{\mathsf{T}\in\mathcal{T}^B}{\text{minimize}}\, \mathcal{D}(\mathsf{T}_\sharp\eta, \nu)$$

has a unique minimizer $\mathsf{T}^\dagger$ *achieving* $\mathcal{D}(\mathsf{T}^\dagger_\sharp\eta, \nu) = 0$ *and* $\mathsf{G}^\dagger(y,\cdot)_\sharp\eta_{\mathcal{U}} = \nu(\cdot|y)$ *for* $\nu_{\mathcal{Y}}$*-a.e.* y*.*

Proof. We begin by showing the existence of a minimizer $\mathsf{T}^\dagger$ which achieves $\mathcal{D}(\mathsf{T}^\dagger_\sharp\eta, \nu) = 0$. Since we assumed $\eta_{\mathcal{Y}}$ has no atoms it follows from the celebrated result of McCann [75, Main Theorem] that there exists a unique map $\mathsf{F}^\dagger\colon \mathcal{Y}\to\mathcal{Y}$ which is the gradient of a convex function and $\mathsf{F}^\dagger_\sharp\eta_{\mathcal{Y}} = \nu_{\mathcal{Y}}$. Thus $(\mathsf{F}^\dagger\times \mathrm{Id})_\sharp\eta = \nu_{\mathcal{Y}}\otimes\eta_{\mathcal{U}}$. Let $\mathsf{G}^\dagger$ denote the unique monotone map from Proposition 2.12 and define $\widetilde{\mathsf{T}}\colon (y,u)\mapsto(y,\mathsf{G}^\dagger(y,u))$. Now observe that $\mathsf{T}^\dagger = \widetilde{\mathsf{T}}\circ(\mathsf{F}^\dagger\times\mathrm{Id})$ satisfies $\mathsf{T}^\dagger_\sharp\eta = \nu$ and belongs to $\mathcal{T}^B$ by construction.

We now verify the uniqueness of $\mathsf{T}^\dagger$. Suppose there exists another map $\mathsf{T}'\in\mathcal{T}^B$, with components F', G' given by gradients of convex functions, and such that $\mathsf{T}'_\sharp\eta = \nu$. By definition the $\mathcal{Y}$ marginal of $\mathsf{T}'_\sharp\eta$ coincides with $\nu_{\mathcal{Y}}$ and so $\mathsf{F}'_\sharp\eta_{\mathcal{Y}} = \nu_{\mathcal{Y}}$ thanks to the product structure of η. It follows from the uniqueness of $\mathsf{F}^\dagger$ that we should have $\mathsf{F}' = \mathsf{F}^\dagger$. On the other hand, Theorem 2.4 implies that $\mathsf{G}'(y,\cdot)_\sharp\eta_{\mathcal{U}} = \nu(\cdot|y)$ for $\nu_{\mathcal{Y}}$-a.e. $y\in\mathcal{Y}$. It follows from Proposition 2.12 that we should have $\mathsf{G}' = \mathsf{G}^\dagger$, which yields the desired result. ∎

Remark 2.14. We observe in subsection 4.3 that using the space $\mathcal{T}^M$ still produces numerical solutions that approach the OT map of Theorem 2.13, meaning that these minimizers over $\mathcal{T}^M$ are "close to" gradients of convex functions. Characterizing this behavior is an interesting direction of future research.

Remark 2.15. We highlight that Theorem 2.13 applies only to the finite-dimensional setting and in particular when $\mathcal{S} = \mathcal{Z}$. Our numerical algorithms in section 3 and some of the experiments in section 4 will extend outside of these settings. Thus, there are still gaps in our theory that pose interesting directions for future research. An extension of Theorem 2.13 to the infinite-dimensional setting will require extending Proposition 2.12, which in turn relies heavily on McCann's result. Existence and uniqueness of Brenier maps in infinite dimensions is a contemporary topic in OT and is only known under certain assumptions on the underlying spaces and on the reference and target measures [6, 34, 65].

3. The monotone GAN framework. This section develops a practical framework for solving problems of the form (1.2), motivated by the analysis of section 2. We primarily focus on settings where we only have access to samples from the reference η and the target ν. Henceforth, we write $\widehat{\eta}^N$, $\widehat{\nu}^N$ to denote empirical approximations to the respective measures with N independent and identically distributed (i.i.d.) samples. That is,

$$\widehat{\eta}^N := \frac{1}{N}\sum_{j=1}^N \delta_{s_j}, \qquad \widehat{\nu}^N := \frac{1}{N}\sum_{j=1}^N \delta_{z_j}, \qquad s_j \overset{\text{iid}}{\sim} \eta, \qquad z_j \overset{\text{iid}}{\sim} \nu.$$

We note that the methodology presented in this section will readily generalize to settings where different number of samples are available from η, ν, but we choose to take N samples from both measures for simplicity of presentation. We propose to approximate (1.2)

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$$\min_{\mathsf{T}\in\mathcal{T}_\theta}\max_{g\in\mathcal{F}_\omega} \mathcal{J}(\mathsf{T}_\sharp\widehat{\eta}^N, \widehat{\nu}^N; g) + \mathcal{R}(\mathsf{T}; \widehat{\eta}^N), \tag{3.1}$$

where $\mathcal{T}_\theta \subset \mathcal{T}$ is a space of block triangular maps of the form (2.4), parameterized by θ; $\mathcal{F}_\omega$ is an appropriate space of functions $g\colon \mathcal{Z} \to \mathcal{R}$ known as *discriminators,* parameterized by ω; and $\mathcal{R}\colon \mathcal{T}_\theta \times \mathbb{P}(\mathcal{Z}) \to \mathbb{R}$ is a regularization functional. The functional $\mathcal{J}\colon \mathbb{P}(\mathcal{Z})\times\mathbb{P}(\mathcal{Z})\times\mathcal{F}_\omega \to \mathbb{R}$ is chosen so that $\max_{g\in\mathcal{F}_\omega}\mathcal{J}(\cdot,\cdot;g)$ approximates a distance measure $\mathcal{D}$ such as an integral probability metric or f-divergence [78, 17]. Our interest in such divergences stems from their successful deployment in large-scale problems in ML, particularly in vision [60]. We will further discuss the choice of $\mathcal{J}$ and the role of the discriminator in subsection 3.3.

Let $\mathsf{T}^\dagger_\theta(w,v) = (\mathsf{F}^\dagger_\theta(w), \mathsf{G}^\dagger_\theta(\mathsf{F}^\dagger_\theta(w), v))$ denote a minimizer of (3.1), suppressing its dependence on $N, \mathcal{F}_\omega$, and $\mathcal{R}$. Our hope is that $\mathsf{G}^\dagger_\theta$ is a good approximation to a true conditioning map $\mathsf{G}^\dagger$, such as the map from Theorem 2.13. We can then approximately sample the conditional measure $\nu(\cdot|y^*)$ for a fixed $y^* \in \mathcal{Y}$ by drawing $v_j \overset{\text{iid}}{\sim} \eta_\mathcal{V}$ and evaluating $u_j = \mathsf{G}^\dagger_\theta(y^*, v_j)$.

In the remainder of this section we outline the details of our conditional simulation framework based on solving (3.1). We discuss neural network parameterizations of T and g in subsection 3.1, followed by our choices of regularization $\mathcal{R}$ in subsection 3.2 and of the objective functional $\mathcal{J}$ in subsection 3.3. We summarize our algorithm in subsection 3.4, followed by a discussion of how our approach can be used to solve likelihood-free Bayesian inference problems in subsection 3.5.

3.1. Neural network parameterizations of $\mathcal{T}_\theta$ and $\mathcal{F}_\omega$. Let $\mathcal{X}, \mathcal{H}$ be separable Banach spaces. Below we define the notion of a neural network mapping $\mathcal{X} \to \mathcal{H}$; our construction is inspired by general families of neural operators such as those found in [68, 69, 67]. Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be a fixed function, henceforth referred to as an *activation function.* We overload our notation by writing $\sigma(\mathbf{v}) = (\sigma(v_1), \dots \sigma(v_k))^\top \in \mathbb{R}^k$ for any vector $\mathbf{v} \in \mathbb{R}^k$. Let us fix an integer $L \geq 1$ (i.e., the depth parameter), the vector of integers $\mathbf{d} = (d_0, d_1, \dots, d_L) \in \mathbb{N}^L$ (i.e., the width parameters), activation functions $\sigma^{(\ell)} : \mathbb{R} \to \mathbb{R}$, matrices $\mathbf{W}^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ (i.e., the weights), vectors $\mathbf{b}^{(\ell)} \in \mathbb{R}^{d_\ell}$ (i.e., the biases), and bounded linear operators $\Psi^I\colon \mathcal{X} \to \mathbb{R}^{d_0}$ and $\Psi^O\colon \mathbb{R}^{d_L} \to \mathcal{H}$. We then say that a map $\mathsf{Q}_\alpha\colon \mathcal{X} \to \mathcal{H}$ is a neural network if it has the form

$$\mathsf{Q}_\alpha(x) = \Psi^O\left(\sigma^{(L)}\left(\mathbf{W}^{(L)}\mathbf{x}^{(L)} + \mathbf{b}^{(L)}\right)\right), \quad \mathbf{x}^{(\ell)} = \sigma^{(\ell)}\left(\mathbf{W}^{(\ell)}\mathbf{x}^{(\ell-1)} + \mathbf{b}^{(\ell)}\right), \quad \mathbf{x}^{(0)} = \Psi^I(x),$$

where we use $\alpha := \{(\mathbf{W}^{(\ell)}, \mathbf{b}^{(\ell)})\}_{\ell=1}^L$ to denote the collection of weights and biases of the neural network Q_α. Furthermore, we refer to the collection of integers $L, d_0, \dots, d_L$ together with activation functions $\sigma^{(1)}, \dots, \sigma^{(L)}$ and the operators Ψ^I, Ψ^O as the *architecture* of Q_α. To this end, we define the spaces of neural networks sharing the same architecture as

$$\mathcal{Q}_\alpha(\mathcal{X}, \mathcal{H}; \mathcal{A}) := \left\{\mathsf{Q}_\alpha : \mathcal{X} \to \mathcal{H} \,|\, \mathsf{Q}_\alpha \text{ has architecture } \mathcal{A} = \{L, \mathbf{d}, \sigma^{(1)}, \dots, \sigma^{(L)}, \Psi^I, \Psi^O\}\right\}.$$

With the above notation, we then consider the spaces for the maps T and the discriminator f given by

$$\begin{aligned}
\mathcal{F}_\omega &= \mathcal{Q}_\omega(\mathcal{Z}; \mathcal{R}; \mathcal{A}_0), \\
\mathcal{T}_\theta &= \Big\{\mathsf{T} : \mathcal{S} \to \mathcal{Z},\, \theta = (\theta_1, \theta_2) \;|\quad \mathsf{T}(w,v) = (\mathsf{F}_{\theta_1}(w), \mathsf{G}_{\theta_2}(\mathsf{F}_{\theta_1}(w), v)), \\
&\qquad \mathsf{F}_{\theta_1} \in \mathcal{Q}_{\theta_1}(\mathcal{W}; \mathcal{Y}; \mathcal{A}_1), \mathsf{G}_{\theta_2} \in \mathcal{Q}_{\theta_2}(\mathcal{Z}; \mathcal{U}; \mathcal{A}_2)\Big\}.
\end{aligned}$$

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

For brevity, the architectures $\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2$ are suppressed in our notation for $\mathcal{F}_\omega, \mathcal{T}_\theta, \mathcal{T}_\theta^B$ and will be identified on a case-by-case basis for the numerical experiments in section 4.

Remark 3.1. We note that our choice of the space $\mathcal{T}_\theta$ is completely generic and does not impose any form of monotonicity on the components $\mathsf{F}_{\theta_1}, \mathsf{G}_{\theta_2}$. We make this choice to have maximal flexibility in the design of architectures and to allow practitioners to utilize existing optimal architectures for the task at hand. Instead, we impose monotonicity using a penalty term that is outlined in subsection 3.2. An alternative approach would be to directly parameterize F_{θ_1}, G_{θ_2} as gradients of (partially) convex functions [7]. Indeed, such parameterizations have already been used for approximation of OT maps [48, 72, 82], but the expressivity of the associated input-convex neural networks and the design of appropriate architectures in high-dimensional examples are not well understood, so we do not utilize these constructions here.

3.2. Monotonicity penalty. In order to make our maps T_θ (approximately) monotone we propose to regularize (3.1) using an average monotonicity penalty. More precisely, let $\mathcal{Z} = \mathcal{Y} \times \mathcal{U}$ as before and suppose that $\mathcal{S} = \mathcal{Z}$, i.e., that T maps $\mathcal{Z} \to \mathcal{Z}$. Then we propose the idealized penalty term

$$\mathcal{R}(\mathsf{T}; \mu) = -\lambda \mathbb{E}_{z\sim\mu} \mathbb{E}_{z'\sim\mu} \langle \mathsf{T}(z) - \mathsf{T}(z'), z - z' \rangle_{\mathcal{Z}}$$

for some positive constant $\lambda > 0$ and any measure $\mu \in \mathbb{P}(\mathcal{Z})$. Including this regularization term in the minimization problem (3.1) in fact encourages T to be increasing (in the prescribed sense) over regions that μ endows with mass. It is natural for us to take $\mu \equiv \eta$ so that the regularization term encourages monotonicity of the map at inputs in the support of the reference distribution. In practice, one can set $\mu \equiv \widehat{\eta}^N$, or alternatively generate a new set of i.i.d. samples from the reference and use those samples to evaluate the penalty term.

It is important to note that while the average monotonicity penalty does not ensure that T is monotone everywhere (not even on the support of μ), numerically we find that this regularization term is sufficient to ensure that T is monotone with high probability. In particular, one can easily compute an empirical approximation to

$$\mathbb{P}_{z\sim\mu, z'\sim\mu}[\langle \mathsf{T}(z) - \mathsf{T}(z'), z - z' \rangle_{\mathcal{Z}} > 0], \tag{3.2}$$

which can be tracked during training as a proxy for the map's monotonicity over $\operatorname{supp}(\mu)$.

3.3. Choosing the functional $\mathcal{J}$. The appropriate choice of the functional $\mathcal{J}$ is a compromise between the computational cost of solving (3.1) and the quality of the minimizer as an approximation to the solution of (2.3). Various popular choices of the divergence $\mathcal{D}$ exist in the literature. Most notably, the forward KL divergence (equivalently, maximum likelihood estimation) is widely used for finding invertible triangular maps and NFs [84], and the reverse KL divergence is used for VI or for training autoencoders [90, 19]. Other possible choices include maximum mean discrepancy [16], Wasserstein distances [11], and f-divergences [81, 114].

Our proposed framework is not tied to a specific choice of $\mathcal{J}$. Our theoretical results suggest that so long as $\mathcal{J}$ yields a good approximation to a divergence, then the resulting algorithm should be capable of approximating the map $\mathsf{G}^\dagger$ well. In our current setup, our main requirement on $\mathcal{J}$ is that it does not involve the pushforward density $\mathsf{G}(y, \cdot)_\sharp \eta_{\mathcal{U}}$, as

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

computing this quantity would require evaluating the inverse of the map $u \mapsto \mathsf{G}(y,u)$ and its Jacobian determinant, which are computationally intensive.

Our choice of $\mathcal{J}$ in (3.1) can be motivated by the family of f-divergences [17]: Let $f\colon \Omega_f \subseteq \mathbb{R} \to \mathbb{R}$ be a continuous, convex function with $f(1)=0$ and let $f^*\colon \Omega_{f^*} \subseteq \mathbb{R} \to \mathbb{R}$ denote its Legendre transform. We then define the f-divergence $\mathcal{D}_f\colon \mathbb{P}(\mathcal{Z}) \times \mathbb{P}(\mathcal{Z}) \to \mathbb{R}$ as

$$\mathcal{D}_f(\mu_1,\mu_2) = \sup_{g\in\mathcal{F}} \int g(z)\mu_1(\mathrm{d}z) - \int f^*(g(z))\mu_2(\mathrm{d}z), \tag{3.3}$$

where $\mathcal{F}$ is a class of (measurable) functions mapping $\mathcal{Z} \to \Omega_{f^*}$, often referred to as the discriminator class in the terminology of generative adversarial models. We note that the standard definition of an f-divergence, as in [4], matches our definition in (3.3) under the assumption μ_1 and μ_2 are equivalent measures, i.e., $\mu_1 \sim \mu_2$.[4] For specific choices of Legendre transforms f^*, we also observe that the objective functional $\max_{g\in\mathcal{F}_\omega} \mathcal{J}(\mathsf{T}_\sharp\widehat{\eta}^N, \widehat{\nu}^N; g)$ in (3.1) can be viewed as an approximation to the divergence $\mathcal{D}_f(\mathsf{T}_\sharp\widehat{\eta}^N, \widehat{\nu}^N)$, where the space $\mathcal{F}$ is replaced by a (possibly parametric) class $\mathcal{F}_\omega$. We now outline some choices of $\mathcal{D}_f$ and the associated functional $\mathcal{J}$ that we used in our experiments in section 4.

- Following [81, Table 2], we can choose $f^*(t) = -\log(2-\exp(t))$, which leads to a generalization of the Jensen–Shannon divergence $\mathcal{D}_f$. Moreover, reparameterizing the discriminator as $g(z) = \log 2v(z)$ for some measurable function[5] $v : \mathcal{Z} \to (0,1)$ yields
$$\begin{aligned}\mathcal{D}_{\mathrm{GAN}}(\mu_1,\mu_2) &:= \sup_{v\in\mathcal{F}} \int \log v(z)\mu_1(\mathrm{d}z) + \int \log(1-v(z))\mu_2(\mathrm{d}z)\\ &=: \sup_{v\in\mathcal{F}} \mathcal{J}_{\mathrm{GAN}}(\mu_1,\mu_2;v),\end{aligned}$$
which is precisely the original GAN loss of [41], up to the additive constant $\log(4)$.
- We may take $f^*(t) = t$ and $\mathcal{F}$ to be the class of Lipschitz-1 functions on $\mathcal{Z}$ to obtain the dual formulation of the Wasserstein-1 distance
$$\mathcal{D}_{\mathrm{W}_1}(\mu_1,\mu_2) := \sup_{g\in\mathrm{Lip}_1} \int g(z)\mu_1(\mathrm{d}z) - \int g(z)\mu_2(\mathrm{d}z).$$
In the case where $\mathcal{Z}$ is a finite-dimensional Euclidean space the Lipschitz-1 constraint on g can be further relaxed to a gradient penalty (GP) to obtain
$$\begin{aligned}\mathcal{D}_{\mathrm{WGP}}(\mu_1,\mu_2) &:= \sup_{g\in C^1(\mathcal{Z})} \int g(z)\mu_1(\mathrm{d}z) - \int g(z)\mu_2(\mathrm{d}z) + \gamma \int (\|\nabla g(z)\| - 1)^2 \mu^*(\mathrm{d}z)\\ &=: \sup_{g\in C^1(\mathcal{Z})} \mathcal{J}_{\mathrm{WGP}}(\mu_1,\mu_2;g),\end{aligned}$$
where $\gamma > 0$ is a penalty coefficient. The measure μ^* is somewhat arbitrary, but the following particular choice yields the Wasserstein-GAN GP loss of [42]:
$$\mu^* := \mathrm{Law}\{z \mid z = \alpha z_1 + (1-\alpha)z_2, \quad \alpha \sim U[0,1],\ z_1 \sim \mu_1,\ z_2 \sim \mu_2\}.$$

[4] This follows by a calculation similar to [81, eq. (4)] with the observation that the continuity assumption on f allows us to use [92, Thm. 14.60] to obtain equality, instead of a lower bound, between an integral probability metric and an f-divergence.

[5] Here, we mean measurable as being with respect to a measure to which both μ_1, μ_2 are absolutely continuous, e.g., $1/2\mu_1 + 1/2\mu_2$.

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

- In some of our low-dimensional examples we shall also use the least squares GAN (LS) functional of [73] due to its simplicity and the fact that it showed good empirical performance. That is,

$$\mathcal{J}_{\mathrm{LS}}(\mu_1,\mu_2;g) := \frac{1}{2}\left[\int (g(z)-a)^2 - (g(z)-b)^2\mu_1(\mathrm{d}z) - \int (g(z)-c)^2\mu_2(\mathrm{d}z)\right],$$

  where a,b,c are scalar constants, which we simply chose as $a=c=1$ and $b=0$. Note that, unlike other loss functions, the LS loss cannot be written as an f-divergence. In practice, this functional is optimized using alternating steps that maximize the last two terms with respect to g, and minimize the first term alone with respect to μ_2.

We reiterate that our above choices of $\mathcal{J}$ are driven by empirical success with numerical experiments in section 4. However, the differences between these losses for our experiments were fairly small, suggesting that the performance of the method is not very sensitive to the choice of $\mathcal{J}$ and in practical applications one may simple choose a $\mathcal{J}$ that is simple and convenient to train. At the same time, the question of the optimal choice of $\mathcal{J}$ and more broadly the divergence $\mathcal{D}$, however, is a contemporary topic in generative modeling and is the subject of intense research [17, 81, 17]. Our goal is not to make a statement on the best choice of these functionals, but rather to focus on the transport methodology for conditional simulation.

3.4. Summary of the algorithm. We now present a summary of the M-GAN training procedure, leading to Algorithm 3.1, which is used in the numerical experiments of section 4. To learn the parameters θ,ω for T, f, we solve (3.1) using an alternating gradient descent procedure that is common for training GANs [41]. This approach repeats the following two steps: (1) update the parameters θ while holding ω fixed; and (2) update the parameters ω while holding θ fixed. Informally, the first step improves the map T so that the pushforward samples are closer to $\widehat{\nu}^N$. The second step updates the discriminator g to better distinguish "real" and "fake" samples from $\widehat{\nu}^N$ and $\mathsf{T}_\sharp\widehat{\eta}^N$, respectively. We note that when using the WGP functional for the image in-painting example of subsection 4.6, θ is updated once for every five updates of ω, as is standard practice for large-scale problems.

For conditional sampling, we are primarily interested in approximating the component function G of the map T, which pushes forward $\eta_{\mathcal{V}}$ to $\nu(\cdot|y)$. Thus, for the experiments below, we choose $\mathsf{F}=\mathrm{Id}$, which implies that $\eta_{\mathcal{W}}=\nu_{\mathcal{Y}}$. We also choose $\mathcal{V}=\mathcal{U}$, but with $\eta_{\mathcal{V}}\equiv\eta_{\mathcal{U}}\neq\nu_{\mathcal{U}}$ in general. To simplify notation, we thus write the product reference measure as $\eta=\nu_{\mathcal{Y}}\otimes\eta_{\mathcal{U}}$, as was done in subsection 2.3.

At each gradient descent step, we replace the expectations in $\mathcal{J}$ with empirical averages over minibatches from the reference measure $\widehat{\eta}^N$ and the target measure $\widehat{\nu}^N$, as in the standard GAN training procedure [41]. In particular, for each update of the parameters θ,ω, two minibatches of size M are sampled: one from the reference marginal $\eta_{\mathcal{U}}$, and one from the training set of the joint target $\widehat{\nu}^N$. We then form the two joint empirical measures $\widehat{\eta}^M$ and $\widehat{\nu}^M$ using the same samples $y_i\sim\widehat{\nu}^N_{\mathcal{Y}}$. For some objective functionals (e.g., LS), we found good empirical performance by also sampling an independent minibatch of size M from the marginal $\widetilde{y}_i\sim\widehat{\nu}^N_{\mathcal{Y}}$ and forming the empirical measure $\widehat{\eta}^M$ from $\{(\widetilde{y}_i,v_i)\}_{i=1}^M$ where $v_i\sim\eta_{\mathcal{U}}$, although this extra sampling step is not strictly necessary for unbiased estimates of the objective. We update the

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

Algorithm 3.1 Outline of the M-GAN training procedure.

Input: Target samples $\{(y_j, u_j)\}_{j=1}^N \overset{\text{iid}}{\sim} \nu$, monotonicity penalty parameter $\lambda > 0$, number of epochs, batch size M
Output: Mapping $\mathsf{G} \in \mathcal{Q}_{\theta_2}$ satisfying $\mathsf{G}(y,\cdot)_\sharp \eta_{\mathcal{U}} \approx \nu(\cdot|y)$ for any $y \in \mathcal{Y}$
for number of epochs **do**
  Sample minibatch of size M from training set $(y_1, u_1), \ldots, (y_M, u_M) \overset{\text{iid}}{\sim} \widehat{\nu}^N$
  Draw M reference samples $v_1, \ldots, v_M \overset{\text{iid}}{\sim} \eta_{\mathcal{U}}$
  Form empirical measures $\widehat{\eta}^M$ and $\widehat{\nu}^M$ from $\{(y_i, v_i)\}_{i=1}^M$ and $\{(y_i, u_i)\}_{i=1}^M$, respectively
  Update θ in G by descending its stochastic gradient $\nabla_\theta \mathcal{J}(\mathsf{T}_\sharp \widehat{\eta}^M, \widehat{\nu}^M; g) + \mathcal{R}(\mathsf{T}; \widehat{\eta}^M)$
  Update ω in g by ascending its stochastic gradient $\nabla_\omega \mathcal{J}(\mathsf{T}_\sharp \widehat{\eta}^M, \widehat{\nu}^M; g)$
end for

parameters for multiple epochs (i.e., passes through the training set), until the evaluations of the functional $\mathcal{J}$ and the penalty $\mathcal{R}$ converge.

3.5. Likelihood-free Bayesian inference. Since conditional simulation is a fundamental task in Bayesian inference, we use this section to illustrate how M-GANs can be used for likelihood-free Bayesian inference. To this end, let $\mathcal{Y}$ and $\mathcal{U}$ denote the data space and parameter space of interest, respectively, and let the conditional measure $\nu(\cdot|u)$ represent a statistical model for the data $y \in \mathcal{Y}$, parameterized by $u \in \mathcal{U}$. Furthermore, let $\nu_0 \in \mathbb{P}(\mathcal{U})$ denote a *prior* measure on u. Then, the goal of Bayesian inference is to characterize the conditional measures $\nu(\cdot|y)$, which is the system of conditionals of the joint measure $\nu(\mathrm{d}y, \mathrm{d}u) = \nu(\mathrm{d}y|u)\nu_0(\mathrm{d}u) \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$, where $\nu(\cdot|u)$ are regarded as transition kernels.

Let us now consider a reference measure $\eta = \eta_{\mathcal{Y}} \otimes \eta_{\mathcal{U}} \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$. Then from any $\mathsf{T}^\dagger$ that is a global minimizer of (2.3) we can extract $\mathsf{G}^\dagger$ and it follows from Theorem 2.4 that

$$\mathsf{G}^\dagger(y,\cdot)_\sharp \eta_{\mathcal{U}} = \nu(\cdot|y) \qquad \text{for } \nu_{\mathcal{Y}}\text{-a.e. } y \in \mathcal{Y}. \tag{3.4}$$

In other words, $\mathsf{G}^\dagger$ completely characterizes the posterior measure for all values of $y \sim \nu_{\mathcal{Y}}$.

We make two key observations about the map. First, the identity (3.4) suggests that in the case of Bayesian inverse problems it is reasonable to choose $\eta_{\mathcal{U}} = \nu_0$, i.e., to choose the prior as the reference measure on $\mathcal{U}$. This choice is motivated by the fact that posterior measures often deviate from the prior (essentially) over a low-dimensional subspace [31, 30, 23]. Hence, one expects that this choice of the reference would result in a map $\mathsf{G}^\dagger$ that is also (essentially) low-dimensional and that captures how the posterior deviates from the prior. Second, similarly to other conditional generative models [83], the map $\mathsf{G}^\dagger$ provides a single function for cheaply sampling from the conditional $\nu(\cdot|y)$ given any realization of the data $y \in \operatorname{supp} \nu_{\mathcal{Y}}$. In comparison to traditional sampling algorithms (e.g., MCMC) that must be repeated for each new realization of y, the process of learning this single map amortizes the cost of inference over the data.

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

To make our approach concrete, we simulate a set of samples $u_j \overset{\text{iid}}{\sim} \nu_0$ and evaluate our forward model to obtain corresponding observations $y_j \overset{\text{iid}}{\sim} \nu(\cdot|u_j)$. The tuples (y_j, u_j) are then draws from the target joint measure ν. We draw an additional set of samples $\tilde{u}_j \sim \nu_0$ so that $(y_j, \tilde{u}_j)$ are draws from the reference distribution $\eta = \nu_{\mathcal{Y}} \otimes \nu_0$. We then apply the M-GAN approach to identify the posterior using the aforementioned empirical samples.

We remark that a likelihood function does not appear in the optimization problems (2.3) or (3.1) and so we only need to evaluate the forward map when generating the samples $y_j \sim \nu(\cdot|u_j)$. In the case of PDE-constrained inverse problems (see, for example, subsection 4.5), simulating the measurements y_j is the most costly step in generating the data for M-GANs. Given that these samples are independent, however, they can be generated in parallel, which is a major advantage over standard MCMC algorithms.

4. Numerical experiments.

We now present a series of experiments that demonstrate the effectiveness of M-GANs in various conditional sampling applications. We emphasize that the goal of these experiments is to highlight the versatility and wide applicability of the M-GAN formulation and block triangular maps in general, rather than focusing on state-of-the-art performance through tuning of network architectures and training recipes. Indeed, our neural networks can easily be replaced with the latest GAN or NF architectures to improve task-specific performance. In subsection 4.1 we demonstrate the importance of the monotonicity constraint for accurate uncertainty quantification on nonlinear regression problems with non-Gaussian noise models. In subsection 4.2 we show that block triangular maps are insensitive to variable ordering, in contrast with strictly triangular maps. Subsection 4.3 shows that the M-GAN framework can recover L^2 OT maps. In subsections 4.4 and 4.5 we present applications to two inverse problems with non-Gaussian posterior measures: parameter inference in coupled ODEs and a Darcy flow model. Finally, subsection 4.6 demonstrates the feasibility of M-GAN in high-dimensional conditional sampling problems arising in imaging. Unless otherwise stated, we take the reference measure to be $\eta = \nu_{\mathcal{Y}} \otimes N(0, I)$; that is, η has the same y marginal as the training data, allowing us to take $\mathsf{F} = \mathrm{Id}$, while the u marginal is a standard Gaussian of the appropriate dimension. Code to reproduce the numerical results is available online at www.github.com/baptistar/MGAN.

4.1. Synthetic examples.

We start with a simple set of synthetic examples where the conditionals $\nu(\cdot|y)$ can be computed explicitly. Consider the following input-to-output maps:

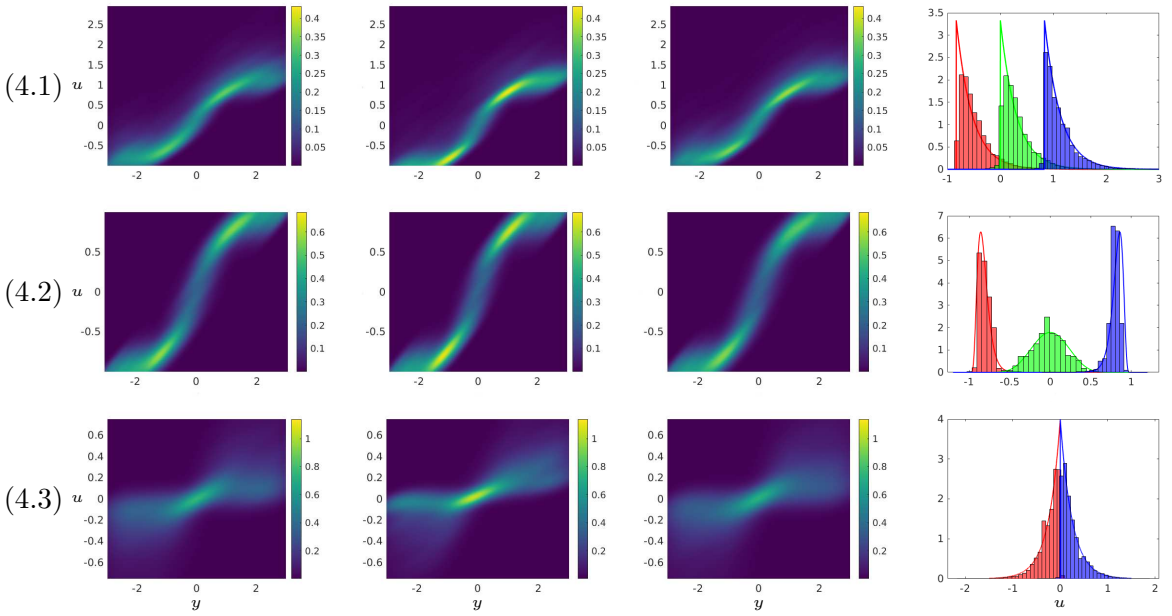
$$u = \tanh(y) + \xi, \qquad \xi \sim \Gamma(1, 0.3), \tag{4.1}$$

$$u = \tanh(y + \xi), \qquad \xi \sim \mathcal{N}(0, 0.05), \tag{4.2}$$

$$u = \xi \tanh(y), \qquad \xi \sim \Gamma(1, 0.3), \tag{4.3}$$

where $y \sim U[-3, 3]$ in all cases. We considered the problem of conditioning u on y and compared M-GAN maps computed via the LS loss functional using $N = 50{,}000$ training samples and with $\lambda = 0$ (i.e., no monotonicity penalty) and $\lambda = 0.01$. We parameterized each map G as a three-layer, fully connected neural network with hidden layer sizes 256/512/128 and leaky ReLU activation functions [71] with parameter $\alpha = 0.2$. We used the same architecture for our discriminator with an additional linear transformation in the final layer to make the

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



(a) ν. (b) $\lambda = 0$. (c) $\lambda = 0.01$. (d) $\nu(\cdot|y)$.

Figure 4.1. *The rows correspond to problems (4.1), (4.2), and (4.3), respectively. The first three columns compare the true joint densities for ν to kernel density estimates (KDEs) of conditional samples from M-GAN with ($\lambda = 0.01$) and without ($\lambda = 0$) the monotonicity penalty. The last column compares histograms of conditional samples from M-GAN with $\lambda = 0.01$ to the true conditional densities (solid lines) for all three problems. The red, green, and blue colors correspond to the distributions for $u|y = -1.1$, $u|y = 0$, and $u|y = 1.1$.*

output one-dimensional. Training was performed using the Adam algorithm [61] with learning rate 2×10^{-4} and parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We used a batch size of $M = 100$ and trained for 300 epochs.

Figures 4.1(a)–(c) compare the true joint densities ν to the M-GAN approximations with and without the monotonicity penalty. We observe a better match between the true density and the M-GAN pushforward $\mathsf{T}_\sharp \eta$ obtained with the monotonicity penalty, particularly in regions of high probability. Figure 4.1(d) compares histograms of conditional samples obtained from M-GAN to the true conditional PDFs, explicitly showing M-GAN's ability to capture the conditionals correctly.

Since this example is two-dimensional, our map parameterization is immediately strictly triangular, and thus we expect M-GAN to approximate the KR rearrangement, as the latter is a global minimizer of (2.3). Figure 4.2 compares the second component function of the true KR rearrangement to the M-GAN map G, with $\lambda = 0.01$. Interestingly, the M-GAN map approximates the KR map very closely, despite not using the explicit KR construction. We note that the pointwise convergence of G to the KR map also implies that the samples generated using G will be close to the target conditionals following the stability theory of [12], which states that for appropriate divergences D (such as the maximum mean discrepancy and the Wasserstein-2 metric) it holds that $D(\mathsf{G}_\sharp \eta_{\mathcal{U}}, \mathsf{G}^\dagger_\sharp \eta_U) \lesssim \|\mathsf{G} - \mathsf{G}^\dagger\|_{L^2_{\eta_{\mathcal{U}}}}$. Thus, the error between the maps controls the error between the pushforward measures.

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

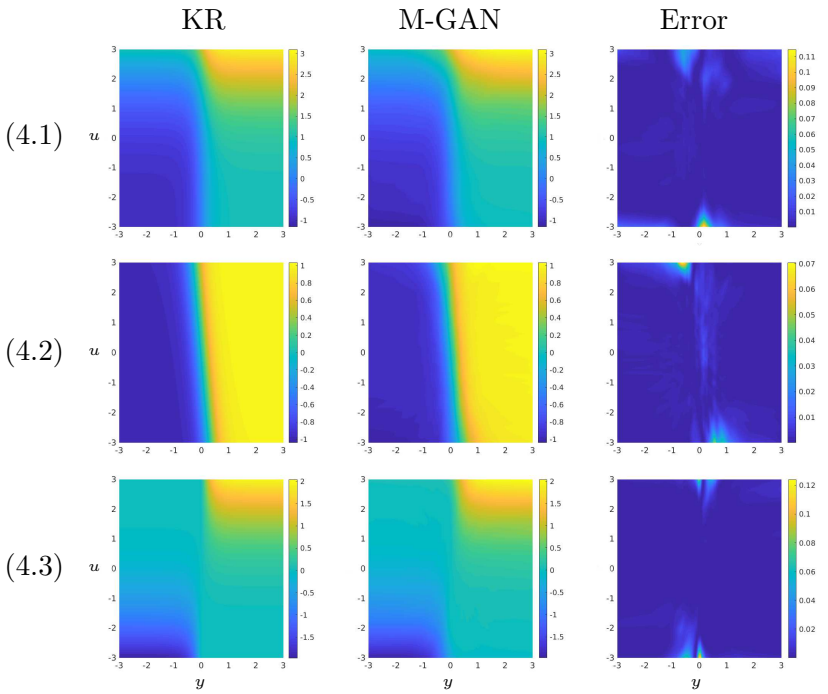


Figure 4.2. *Each row corresponds to the problems* (4.1), (4.2), *and* (4.3), *respectively. The first column shows the (second component function of) the true KR rearrangement, the second column shows the M-GAN map* G, *and the last column shows the absolute pointwise error between the two.*

4.2. Insensitivity to variable ordering. We now illustrate a benefit of using nontriangular maps, rather than triangular maps such as the KR rearrangement. Consider the random vector $u = (u_1, u_2)$ with $u_1 \sim \mathcal{N}(0,1)$ and $u_2|u_1 \sim \mathcal{N}(u_1^2 + 1, 0.5^2)$. For simplicity, we omit the conditioning variables y in this example. The bivariate distribution of u can be represented exactly as the pushforward of $\eta_{\mathcal{V}} = \mathcal{N}(0, I_2)$ by the map $\mathsf{T}(v) = [v_1; v_1^2 + 1 + 0.5v_2]$. Hence, T is easily approximated by a triangular map of this form. When the ordering of u_1 and u_2 are reversed, however—i.e., when the first component of T must represent the marginal of u_2 instead of u_1—the triangular map is more challenging to approximate. To resolve this issue, one common approach is to compose many maps to define an expressive NF; see [85] for a similar application. We demonstrate instead that by using a nontriangular parameterization (which would become block triangular when there are conditioning variables), we can avoid issues pertaining to the ordering of the u variables and achieve a more robust map in practice.

We use $N = 10^4$ training samples, $\lambda = 0.01$, and the LS loss function to train an M-GAN with either triangular or nontriangular structure. We use three-layer fully connected neural networks with hidden layer sizes 32/64/32 for the nontriangular maps and neural networks with hidden layer sizes 22/46/22 for each component of the triangular map. In total, the nontriangular and triangular maps have about the same number of parameters. Both M-GAN maps are trained using the same optimization setup as in subsection 4.1.

Figure 4.3 compares the samples generated by the triangular and nontriangular maps to the true density of u. We observe that the nontriangular map is able to capture the target

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

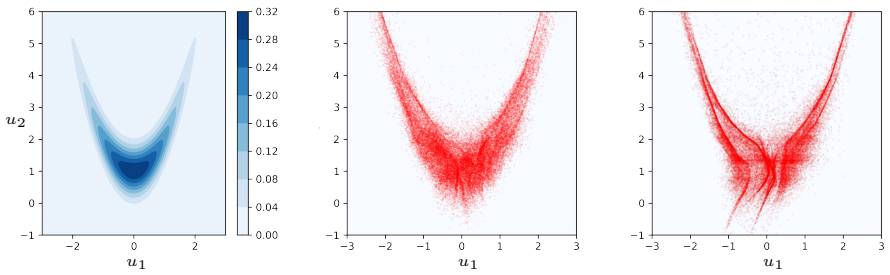


Figure 4.3. *(Left) The true density of (u_1, u_2) considered in subsection 4.2. (Middle) Samples generated by M-GAN using a nontriangular map with the reverse ordering of the variables. (Right) Samples generated by M-GAN using a triangular map, also with reverse ordering.*

Table 4.1

KL divergence errors for nontriangular and triangular M-GANs computed using $N = 10^4$ training samples. The approximate densities are estimated using KDE by generating 5×10^4 samples and using an optimal bandwidth parameter that is chosen using fivefold cross-validation. The KL divergence is evaluated using an average of 10^4 independent test samples and is reported with its 95% standard error in parentheses.

	Nontriangular	Triangular
Favorable order (u_1, u_2)	0.056 (0.003)	0.039 (0.002)
Reverse order (u_2, u_1)	0.058 (0.002)	0.102 (0.004)

density with an unfavorable ordering of the variables, unlike the triangular map. Table 4.1 reports the KL divergence between the true and approximated distributions for both variable orderings. The nontriangular map provides essentially the same performance independent of ordering, while the performance of the triangular map improves or degrades significantly depending on the ordering. This suggests that nontriangular maps are less sensitive to the variable ordering, a major advantage of M-GANs in comparison to autoregressive models where it is necessary to specify a variable ordering in advance.

4.3. Approximation of OT maps. Now we show how the M-GAN framework using an average monotonicity penalty recovers the transport map G that minimizes the L^2 transport cost $\mathbb{E}_{(y,v)\sim\eta}\|v - \mathsf{G}(y,v)\|^2$, i.e., the conditional Brenier map of Proposition 2.12. We consider a multivariate Gaussian distribution ν with $\mathcal{Y} = \mathbb{R}$ and $\mathcal{U} = \mathbb{R}^5$. The marginal distribution of u is chosen to be $\mathcal{N}(m_u, \Sigma_u)$ where the mean and covariance are randomly sampled as $m_u \sim \mathcal{N}(0, \mathrm{I}_5)$ and $\Sigma_u = UU^\top$ for orthonormal column vectors $U \in \mathcal{R}^{5\times 5}$ (from the QR decomposition of a matrix with standard Gaussian entries) and fixed for this experiment. The measurement y is given by $y = u_4 + \xi$ where $\xi \sim \mathcal{N}(0,1)$. By Proposition 2.12, the monotone transport map pushing forward a standard Gaussian reference $\eta_{\mathcal{V}} = \mathcal{N}(0, \mathrm{I}_5)$ to the conditionals $\nu(\cdot|y)$ is unique among all maps that are gradients of a convex function. In this Gaussian case, the optimal map is given by $\mathsf{G}^\dagger(y,v) := m_{u|y}(y) + \Sigma_{u|y}^{1/2} v$; see [24, Example 2.1].

Given $N = 10^4$ training samples from ν, we learn the M-GAN map G using the WGP loss with GP $\gamma = 1$ and three increasing values of the monotonicity penalty λ. We parameterize the maps and the discriminators using three-layer, fully connected neural networks

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

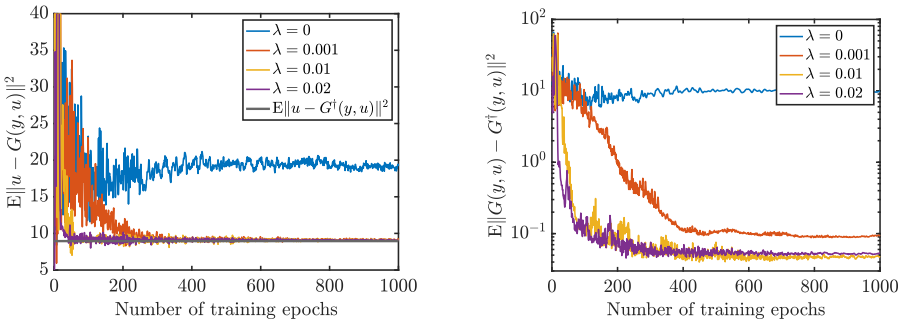


Figure 4.4. *(Left) The transport cost* $\mathbb{E}_{(y,v)\sim\eta}\|v - \mathsf{G}(y,v)\|^2$ *associated with M-GAN maps* G *converges to the* minimal *transport cost, achieved by the optimal map* $\mathsf{G}^\dagger$*, when increasing the monotonicity penalty* λ*. (Right) The maps themselves converge in the* L^2_η *sense to the optimal map* $\mathsf{G}^\dagger$ *when increasing* λ*.*

with hidden layer sizes 64/64/64. While affine maps (in v and y) are sufficient to represent the Gaussian conditionals in this example, our goal is to demonstrate the convergence of the M-GAN training procedure to the conditional Brenier map over the large space of nonlinear functions described in subsection 3.1. We train using the Adam algorithm as in subsection 4.1, with a batch size of $M = 1000$ and a scheduled learning rate that decays by 0.995 starting from 4×10^{-3}, over 1000 epochs.

Figure 4.4 plots the transport cost $\mathbb{E}_{(y,v)\sim\eta}\|v - \mathsf{G}(y,v)\|^2$ for the estimated maps G and the expected squared error between the estimated maps and the optimal map $\mathsf{G}^\dagger$. We observe that maps found with the average monotonicity penalty term indeed converge to the *optimal map* of Proposition 2.12 and similarly that the associated transport cost converges to the OT cost, when increasing the penalty λ from 0 to 0.01. We further observe that choosing λ to be too large leads to diminishing returns and higher errors (or even divergence in the optimization) as is customary in regularized optimization due to excessive bias. It is also important to note that choosing an appropriate value of λ not only improves the accuracy of the map but also improves the rate of convergence of both the map and the loss toward their optimal counterparts. Of course, an alternative approach could involve replacing the λ-dependent monotonicity penalty with the constraint that T lie in $\mathcal{T}^B$; this involves some practical difficulties, as described in Remark 3.1. In the Gaussian case one could write T as the gradient of a quadratic function as in [105, 3], for example.

4.4. Inference of ODE parameters. Next, we use the MGAN framework to infer the parameters in a Lotka–Volterra population model, which is a common benchmark for likelihood-free inference [70]. This model describes the populations of interacting species, such as predators and prey, using nonlinear coupled ODEs where the rates of change of the two populations depend on four parameters $u = (\alpha, \beta, \gamma, \delta) \in \mathcal{R}^4$. Our goal is to infer these parameters given noisy observations of the populations of predators and prey, i.e., the system states, at select times. The states $p(t) \in \mathbb{R}^2_+$ evolve according to the coupled ODEs

$$\frac{dp_1}{dt} = \alpha p_1(t) - \beta p_1(t) p_2(t),$$
$$\frac{dp_2}{dt} = -\gamma p_2(t) + \delta p_1(t) p_2(t),$$

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

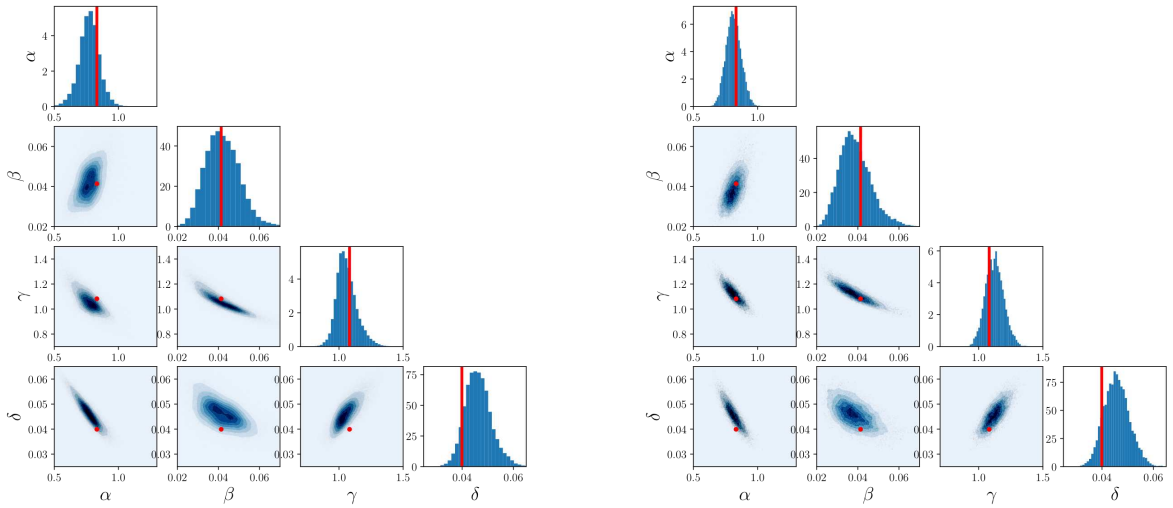


Figure 4.5. *Posterior samples of the parameters in the deterministic Lotka–Volterra model using (left) the M-GAN framework, and (right) the adaptive Metropolis MCMC algorithm.*

with the initial condition $p(0) = (30, 1)$. We simulate the ODEs for $T = 20$ time units and collect noisy observations of the state every $\Delta t_{\text{obs}} = 2$ time units. The observations are corrupted with log-normal noise, i.e., $\log y_k \sim \mathcal{N}(\log p(k\Delta t_{\text{obs}}), \sigma^2 \mathrm{I}_2)$ for $k = 1, \ldots, 9$, with standard deviation $\sigma = 0.01$. For inference, we use an independent log-normal prior distribution for the parameters given by $\log u \sim \mathcal{N}(m_u, 0.5\mathrm{I}_4)$ with $m_u = (-0.125, -3, -0.125, -3)$. Figure 4.6 displays the states $p(t)$ (solid line) for the parameter $u^* = (0.92, 0.05, 1.50, 0.02)$ and an observation $y^* \in \mathbb{R}^{18}$ drawn from the conditional distribution $\nu(\cdot|u^*)$.

We then sample from the posterior density for $u|y = y^*$ given $N = 10^5$ training samples from ν using both M-GAN and an MCMC algorithm. First, we train an M-GAN network with the WGP loss using the monotonicity penalty $\lambda = 0.1$ and the GP $\gamma = 1$. For this example we used three-layer, fully connected neural networks with hidden layer sizes 128/256/512 for the map and hidden layer sizes 512/256/128 for the discriminator. We used the Adam optimizer with the same parameters as in subsection 4.1 and trained for 400 epochs.

Figure 4.5 displays 100,000 parameter samples from M-GAN, i.e., $\mathsf{G}(y^*, u^i)$ for $u^i \sim \mathcal{N}(0, \mathrm{I}_4)$ after learning the map G, and from an adaptive Metropolis MCMC sampler, respectively. We observe similar one- and two-dimensional marginal distributions using both methods. The true parameter u^* that generated the data (denoted in red) is contained in the bulk of the posterior distributions and appears like a representative sample. Last, we integrate the ODEs for sample realizations of the posterior parameters to sample from the predictive distribution for the states $p(t)$. The dashed lines in Figure 4.6 plot 10 posterior predictive samples for both M-GAN and MCMC. We observe that samples from both methods concentrate around the true states and that the predictions from M-GAN have similar spread to MCMC (i.e., the ground truth), especially at earlier times.

4.5. Darcy flow Bayesian inverse problem. We now consider a benchmark inverse problem from subsurface flow modeling [49] and electrical impedance tomography [56] whose forward model is given by the partial differential equation (PDE)

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

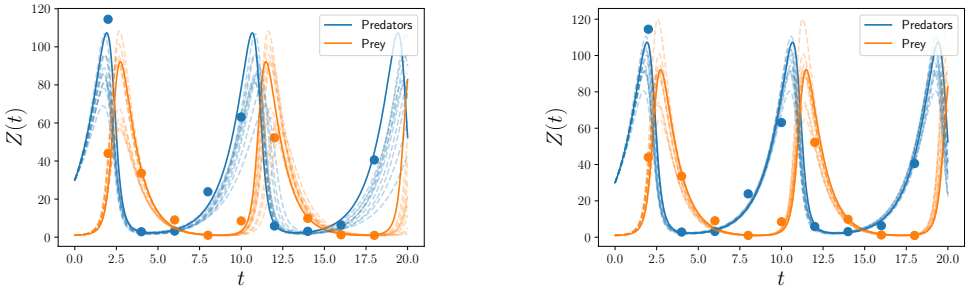
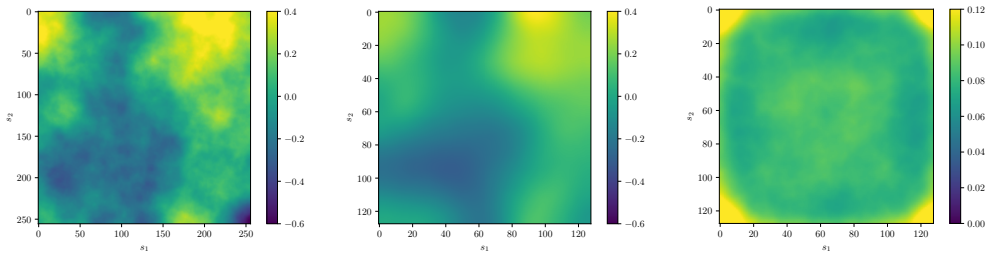
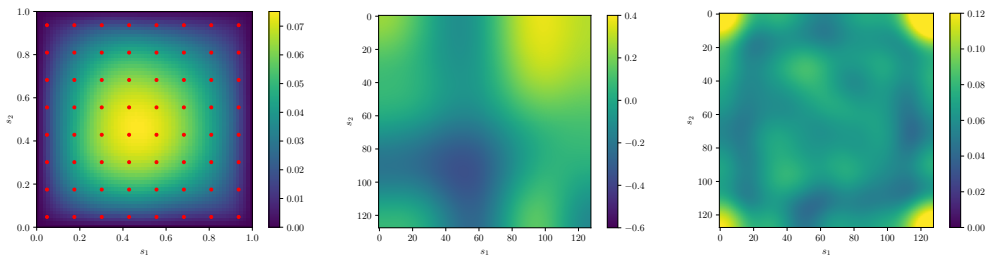


Figure 4.6. *Posterior predictive samples for the states $p(t)$ given 10 posterior samples from (left) the M-GAN map and (right) the MCMC algorithm.*



(a) True field u. (c) pCN mean. (e) pCN std.



(b) Pressure field p. (d) M-GAN mean. (f) M-GAN std.

Figure 4.7. *A comparison of posterior samples for the Darcy flow inverse problem using M-GANs and pCN.* (a) *A realization of the Gaussian random field u.* (b) *The solution to the PDE with the measurement locations denoted in red.* (c) *and* (d) *The mean of the posterior samples from M-GAN and pCN, respectively.* (e) *and* (f) *The standard deviation of the posterior samples from M-GAN and pCN, respectively.*

$$
\begin{aligned}
-\nabla \cdot (a(s)\nabla p(s)) &= 1, \quad s \in (0,1)^2, \\
p(s) &= 0, \quad s \in \partial(0,1)^2.
\end{aligned} \tag{4.4}
$$

We interpret $p(s)$ as the pressure field of subsurface flow in a reservoir with permeability coefficient $a(s) \in \mathbb{R}_+$ under constant forcing. We further introduce a log-normal random field for the permeability given by $u = \log(a) \sim \mathcal{N}(0, (-\Delta + 9I)^{-2})$ where Δ is the Laplacian operator with zero Neumann boundary conditions. The inverse problem is to recover the random field u given noisy measurements y of the pressure at 64 regularly spaced locations, i.e., $y = (p(s_1), \ldots, p(s_{64})) + \gamma$ where $\gamma \sim \mathcal{N}(0, 10^{-6} I_{64})$. Figures 4.7(a) and (d) plot a realization

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

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

of u along with the solution to the PDE on a grid of size 256×256 as well as the location of pressure measurements $s_1, \ldots, s_{64}$.

To recover the permeability a, we train an M-GAN following the discussion of subsection 3.5. We sample a training set of size $N = 10^5$ from ν that was obtained using the following recipe: draw $u \sim \mathcal{N}(0, (-\Delta + 9I)^{-2})$ and set $a = \exp(u)$; solve the PDE using finite differences to obtain p; use spline interpolation to simulate a set of measurements y at the observation locations. The M-GAN map was trained using the WGP loss functional with the GP $\gamma = 1$ and the monotonicity penalty $\lambda = 0.01$. To make our network architectures consistent in the continuum limit, and hence mesh independent, we use PCA projections to reduce the dimension of the u samples at the input to the networks, i.e., the Ψ^I operator in subsection 3.1 is taken to be the PCA projection of the random field u onto its first 25 PCA modes, which capture 99.98% of the total prior variation (as measured by the trace of the prior covariance). To this end, the G component of our M-GAN map takes inputs in $\mathbb{R}^{64} \times \mathbb{R}^{25}$ (64 for y and 25 for the leading PCA modes of u) and outputs a vector of PCA modes of u in $\mathbb{R}^{25}$ which can then be lifted to a random field by taking Ψ^O to be the PCA reconstruction map. In summary, our map G will condition the first 25 PCA coefficients of u on observations of the data y. In this experiment we use the same network architectures and optimizer as in subsection 4.4 and train for 500 epochs.

We used 5×10^4 prior samples in order to compute the PCA modes of u and generated a fixed realization y^* for a single draw of the field u, that is taken to be the ground truth. To avoid any inverse crimes [57] we generated the data y^* using a mesh that was twice as fine as the mesh used to generate the training data. Figures 4.7(b), (e), (c), and (f) compare the posterior mean and the standard deviation for the field u obtained by M-GANs with the preconditioned Crank–Nicolson (pCN) MCMC algorithm [26], which is regarded as the gold standard solution. We tuned the pCN step size to achieve an acceptance rate between 20% and 40% after burn-in and used 10^6 samples to compute the mean and standard deviations. We observe good agreement between the M-GAN mean and pCN while the standard deviation appears to have been slightly underestimated by M-GAN, a feature that is common with prior-based dimension reduction techniques. Note that a more conservative estimate for the standard deviation can be obtained by sampling the trailing PCA coefficients from their prior distribution, i.e., without conditioning on the data, and combining these with the M-GAN posterior samples for the leading PCA coefficients [23, 31].

This experiment not only demonstrates the feasibility of M-GANs for likelihood-free inference on function spaces, but also suggests that M-GANs can potentially lead to improved performance for PDE inverse problems: Our maps were computed using only 10^5 PDE solves while pCN required 10^6 samples to compute a stable estimate of the standard deviation. Moreover, the latter would have to be rerun for any new realization of the data y, whereas the M-GAN map can be applied, without additional training, to any new realization of y. Furthermore, the M-GAN training set can be generated fully in parallel since its samples are independent, unlike MCMC that requires sequential PDE solves for each accept/reject step.

4.6. Probabilistic image in-painting. For our final set of experiments, we consider the "in-painting" problem of reconstructing an image after a portion of it has been removed. We view this problem in our general probabilistic setting, as image-to-image regression where the

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

Figure 4.8. *Example in-paintings using M-GAN for the CelebA test set. The first column depicts the ground truth image, the second column shows the observed image y^*, while the next three columns are random samples from the conditional distribution for $u|y^*$. The last two columns show the pointwise means and variances for the intensities of the conditional samples generated by the M-GAN map.*

input/measurement y is the incomplete image and the output u is the in-painting. The conditional distribution $\nu(\cdot|y)$ thus quantifies uncertainty in the reconstruction, with its samples being understood as candidate in-paintings. We consider the CelebA dataset consisting of $64 \times 64 \times 3$ RGB images of celebrity faces (converted to a standard size using bicubic interpolation). The input $y \in \mathbb{R}^{32\times64\times3}$ consists of the top half of each image, and the output $u \in \mathbb{R}^{32\times64\times3}$ consists of the bottom half.

We trained an M-GAN on the training set of $N = 162\,770$ images using the WGP loss functional with the monotonicity penalty $\lambda = 10^{-4}$ and GP $\gamma = 1$. We also added independent Gaussian white noise with standard deviation 0.05 to corrupt each image in the training set, as in [60]. We chose our reference measure to be $\eta = \nu_{\mathcal{Y}} \otimes \mathcal{N}(0, \mathrm{I}_{100})$, i.e., the $\mathcal{Y}$ marginal coincides with that of the training data, while the latent space for the $\mathcal{U}$ variable is assumed to be $\mathbb{R}^{100}$ equipped with standard Gaussian measure. Hence, the input and output spaces of T do not match in this example, in contrast with our previous experiments. As for the architectures, we used the convolutional architectures introduced in [87] with suitable modifications for our input and output dimensions. We used the same training/optimization setup as in subsection 4.1.

Figure 4.8 shows conditional samples of image in-paintings for the CelebA test set, together with the conditional mean and variance of the pixelwise image intensities. We note the variability among the M-GAN samples, producing different smiles, hair styles, jawlines, outfits, and backgrounds—as one should expect from a probabilistic in-painting method. We also computed a FID score of approximately 35 for the M-GAN map in this example. We emphasize, however, that while FID is a common metric for photorealism, it fails to capture accuracy in characterizing the conditional distributions. For example, we noticed that maps whose range collapses conditionally onto a single point and as result sample the same in-painting $\mathsf{G}(y, w)$ for any realization of $w \sim \mathcal{N}(0, I_{100})$ can still obtain similarly good FID

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

scores, while failing to capture the true conditional distribution. To the best of our knowledge, distributional in-painting on the CelebA dataset has not been explored in the literature, and thus we cannot compare the FID of our result to others. The closest to the state of the art is a FID of 30 reported in [109] for the CelebA-HQ dataset.

5. Conclusions. We have developed M-GAN, a transport-based approach for conditional generative modeling and likelihood-free (simulation-based) inference. Our approach seeks a block triangular transport map that pushes forward a chosen reference measure η to a target measure ν, defined on the joint space of the parameter and data. Under very mild assumptions, essentially that the reference measure has an appropriate product structure, we show that this construction produces a component transport map that captures the conditional measures $\nu(\cdot|y)$ of the target, and that this map enables direct conditional sampling. We propose an adversarial training procedure to learn such a map, incorporating a monotonicity penalty that drives the solution of the optimization problem toward the unique *conditional* OT map minimizing an L^2 transport cost.

Our numerical experiments demonstrate the effectiveness and versatility of M-GANs in applications ranging from parameter inference and inverse problems to imaging, all tackled in an entirely data-driven and likelihood-free setting. In most of our examples we compared M-GAN to MCMC as our gold standard for conditional sampling. This raises the following question**:** *Are M-GAN and, more broadly, any transport-based sampling method better than MCMC and in what sense?* Our numerical results paint an interesting and nuanced picture in response to this question that warrants careful study in the future. The entire training of M-GAN can be performed offline with very fast evaluations at the inference stage, i.e., conditional samples can be generated using M-GAN by simply evaluating the network, which is often orders of magnitude faster than MCMC algorithms. For example, in the case of the Darcy flow example in subsection 4.5, MCMC takes multiple hours to converge on a personal computer. On the other hand, to achieve high accuracy with M-GAN, one may need to employ large networks with specialized architectures, which require long training times and so the learning cost may be an important factor. Finally, M-GAN requires the generation of large amounts of training data which can be costly for complex models such as PDEs. The sample-generation procedure, however, is highly parallelizable and the cost is amortized when solving multiple inverse problems for different values of the conditioning variables, unlike MCMC.

In future research, the interplay between the quality of an M-GAN map obtained by solving the practical optimization problem (3.1) and the accuracy of the derived conditionals warrants theoretical investigation. Relatedly, approximation results characterizing the expressiveness of parametric classes of block triangular maps would be of great interest. It would also be useful to extend the links to OT described here to infinite-dimensional function spaces, as such a connection would be pertinent to inverse problems.

REFERENCES

[1] J. Adler and O. Öktem, *Deep Bayesian Inversion*, preprint, arXiv:1811.05910, 2018.
[2] S. Agapiou, M. Burger, M. Dashti, and T. Helin, *Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems*, Inverse Problems, 34 (2018), 045002.

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

[3] M. Al-Jarrah, B. Hosseini, and A. Taghvaei, *Optimal transport particle filters*, in 62nd IEEE Conference on Decision and Control, Marina Bay Sands, Singapore, 2023.
[4] S. M. Ali and S. D. Silvey, *A general class of coefficients of divergence of one distribution from another*, J. R. Stat. Soc. Ser. B. Methodol., 28 (1966), pp. 131–142.
[5] L. Ambrogioni, U. Güçlü, M. A. van Gerven, and E. Maris, *The Kernel Mixture Network: A Nonparametric Method for Conditional Density Estimation of Continuous Random Variables*, preprint, arXiv:1705.07111, 2017.
[6] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*, Springer, New York, 2005.
[7] B. Amos, L. Xu, and J. Z. Kolter, *Input convex neural networks*, in Proceedings of the International Conference on Machine Learning, PMLR, 2017, pp. 146–155.
[8] M. Arbel and A. Gretton, *Kernel conditional exponential family*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, 2018, pp. 1337–1346.
[9] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe, *Guided Image Generation with Conditional Invertible Neural Networks*, preprint, arXiv:1907.02392, 2019.
[10] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in Proceedings of the International Conference on Machine Learning, 2017, pp. 214–223.
[11] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in Proceedings of the International Conference on Machine Learning, PMLR, 2017, pp. 214–223.
[12] R. Baptista, B. Hosseini, N. B. Kovachki, Y. M. Marzouk, and A. Sagiv, *An Approximation Theory Framework for Measure-Transport Sampling Algorithms*, preprint, arXiv:2302.13965, 2023.
[13] R. Baptista, O. Zahm, and Y. Marzouk, *On the representation and learning of monotone triangular transport maps*, Found. Comput. Math., (2023), https://doi.org/10.1007/s10208-023-09630-x.
[14] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann, *Conditional Image Generation with Score-Based Diffusion Models*, preprint, arXiv:2111.13606, 2021.
[15] M. I. Belghazi, M. Oquab, Y. Lecun, and D. Lopez-Paz, *Learning about an exponential amount of conditional distributions*, in Advances in Neural Information Processing Systems, 2019.
[16] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, *Demystifying MMD GANs*, in Proceedings of the International Conference on Learning Representations (ICLR), 2018.
[17] J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and L. Rey-Bellet, (f,γ)*-divergences: Interpolating between* f*-divergences and integral probability metrics*, J. Mach. Learn. Res., 23 (2022), pp. 1–70.
[18] C. M. Bishop, *Mixture Density Networks*, Technical report NCRG/94/004, Department of Computer Science and Applied Mathematics, Aston University, 1994.
[19] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, *Variational inference: A review for statisticians*, J. Amer. Statist. Assoc., 112 (2017), pp. 859–877.
[20] V. I. Bogachev, *Measure Theory*, Vol. 2, Springer, New York, 2007.
[21] V. I. Bogachev and A. V. Kolesnikov, *Nonlinear transformations of convex measures*, Theory Probab. Appl., 50 (2006), pp. 34–52.
[22] V. I. Bogachev, A. V. Kolesnikov, and K. V. Medvedev, *Triangular transformations of measures*, Sb. Mat., 196 (2005), pp. 309–335.
[23] M. Brennan, D. Bigoni, O. Zahm, A. Spantini, and Y. Marzouk, *Greedy inference with structure-exploiting lazy maps*, in Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 8330–8342.
[24] G. Carlier, V. Chernozhukov, and A. Galichon, *Vector quantile regression: An optimal transport approach*, Ann. Statist., 44 (2016), pp. 1165–1192.
[25] G. Carlier, A. Galichon, and F. Santambrogio, *From Knothe's transport to Brenier's map and a continuation method for optimal transport*, SIAM J. Math. Anal., 41 (2010), pp. 2554–2576.
[26] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC methods for functions: Modifying old algorithms to make them faster*, Statist. Sci., 28 (2013), pp. 424–446.
[27] K. Cranmer, J. Brehmer, and G. Louppe, *The frontier of simulation-based inference*, Proc. Natl. Acad. Sci. USA, 117 (2020), pp. 30055–30062.
[28] I. Crimaldi and L. Pratelli, *Convergence results for conditional expectations*, Bernoulli, 11 (2005), pp. 737–745.

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

[29] T. Cui, S. Dolgov, and O. Zahm, *Scalable conditional deep inverse Rosenblatt transports using tensor-trains and gradient-based dimension reduction*, J. Comput. Phys., 485 (2023), 112103.
[30] T. Cui, K. J. Law, and Y. M. Marzouk, *Dimension-independent likelihood-informed MCMC*, J. Comput. Phys., 304 (2016), pp. 109–137.
[31] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini, *Likelihood-informed dimension reduction for nonlinear inverse problems*, Inverse Problems, 30 (2014), 114015.
[32] C. Doersch, *Tutorial on Variational Autoencoders*, preprint, arXiv:1606.05908, 2016.
[33] T. A. El Moselhy and Y. M. Marzouk, *Bayesian inference with optimal maps*, J. Comput. Phys., 231 (2012), pp. 7815–7850.
[34] D. Feyel and A. S. Üstünel, *Monge-Kantorovitch measure transportation and Monge-Ampere equation on Wiener space*, Probab. Theory Related Fields, 128 (2004), pp. 347–385.
[35] C. W. Fox and S. J. Roberts, *A tutorial on variational Bayesian inference*, Artif. Intell. Rev., 38 (2012), pp. 85–95.
[36] M. Gabrié, G. M. Rotskoff, and E. Vanden-Eijnden, *Adaptive Monte Carlo augmented with normalizing flows*, Proc. Natl. Acad. Sci. USA, 119 (2022), e2109420119.
[37] A. Genevay, M. Cuturi, G. Peyré, and F. Bach, *Stochastic optimization for large-scale optimal transport*, in Advances in Neural Information Processing Systems, Vol. 29, 2016.
[38] S. Gershman and N. Goodman, *Amortized inference in probabilistic reasoning*, in Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 36, 2014.
[39] E. M. Goggin, *Convergence in distribution of conditional expectations*, Ann. Probab., 22 (1994), pp. 1097–1114.
[40] I. Goodfellow, *NIPS* 2016 *Tutorial: Generative Adversarial Networks*, preprint, arXiv:1701.00160, 2016.
[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in Neural Information Processing Systems, 2014.
[42] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, *Improved training of Wasserstein GANs*, in Proceedings of the International Conference on Neural Information Processing Systems, 2017.
[43] M. Hairer, A. M. Stuart, and S. J. Vollmer, *Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions*, Ann. Appl. Probab., 24 (2014), pp. 2455–2490.
[44] M. Hassanaly, A. Glaws, K. Stengel, and R. N. King, *Adversarial sampling of unknown and high-dimensional conditional distributions*, J. Comput. Phys., 450 (2022), 110853.
[45] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, 2009.
[46] B. Hosseini, *Two Metropolis–Hastings algorithms for posterior measures with non-Gaussian priors in infinite dimensions*, SIAM/ASA J. Uncertain. Quantif., 7 (2019), pp. 1185–1223.
[47] B. Hosseini and J. E. Johndrow, *Spectral Gaps and Error Estimates for Infinite-Dimensional Metropolis-Hastings with Non-Gaussian Priors*, preprint, arXiv:1810.00297, 2018.
[48] C.-W. Huang, R. T. Chen, C. Tsirigotis, and A. Courville, *Convex potential flows: Universal probability distributions with optimal transport and convex optimization*, in Proceedings of the International Conference on Learning Representations, 2021.
[49] M. A. Iglesias, K. Lin, and A. M. Stuart, *Well-posed Bayesian geometric inverse problems arising in subsurface flow*, Inverse Problems, 30 (2014), 114001.
[50] N. J. Irons, M. Scetbon, S. Pal, and Z. Harchaoui, *Triangular flows for generative modeling: Statistical consistency, smoothness classes, and fast rates*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 10161–10195.
[51] O. Ivanov, M. Figurnov, and D. Vetrov, *Variational autoencoder with arbitrary conditioning*, in Proceedings of the International Conference on Learning Representations (ICLR), 2019.
[52] P. Jaini, I. Kobyzev, Y. Yu, and M. Brubaker, *Tails of Lipschitz triangular flows*, in Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 4673–4681.
[53] P. Jaini, K. A. Selby, and Y. Yu, *Sum-of-squares polynomial flow*, in Proceedings of the International Conference on Machine Learning, PMLR, 2019, pp. 3009–3018.

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

[54] T. Jebara, *Machine Learning: Discriminative and Generative*, Internat. Ser. Engrg. Comput. Sci. 755, Springer, New York, 2012.
[55] S. I. Kabanikhin, *Inverse and Ill-Posed Problems: Theory and Applications*, De Gruyter, Berlin, 2011.
[56] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, New York, 2005.
[57] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Appl. Math. Sci. 160, Springer, New York, 2006.
[58] O. Kallenberg, *Foundations of Modern Probability*, Probab. Appl. (N. Y.), Springer, New York, 2006.
[59] K. Kan, F.-X. Aubet, T. Januschowski, Y. Park, K. Benidis, L. Ruthotto, and J. Gasthaus, *Multivariate quantile function forecaster*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 10603–10621.
[60] T. Karras, S. Laine, and T. Aila, *A style-based generator architecture for generative adversarial networks*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.
[61] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in Proceedings of the 3rd International Conference on Learning Representations, Y. Bengio and Y. LeCun, eds., 2015.
[62] D. P. Kingma and M. Welling, *Auto-encoding variational Bayes*, in Proceedings of the 2nd International Conference on Learning Representations, 2014.
[63] H. Knothe, *Contributions to the theory of convex bodies*, Michigan Math. J., 4 (1957), pp. 39–52.
[64] I. Kobyzev, S. Prince, and M. Brubaker, *Normalizing flows: An introduction and review of current methods*, IEEE Trans. Pattern Anal. Mach. Intell., 43 (2021), pp. 3964–3979.
[65] A. V. Kolesnikov and M. Röckner, *On continuity equations in infinite dimensions with non-Gaussian reference measure*, J. Funct. Anal., 266 (2014), pp. 4490–4537.
[66] A. Korotin, L. Li, A. Genevay, J. M. Solomon, A. Filippov, and E. Burnaev, *Do neural optimal transport solvers work? A continuous Wasserstein-2 benchmark*, in Advances in Neural Information Processing Systems, Vol. 34, 2021.
[67] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar, *Neural operator: Learning maps between function spaces with applications to PDEs*, J. Mach. Learn. Res., 24 (2023), pp. 1–97.
[68] Z. Li, N. B. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, *Fourier neural operator for parametric partial differential equations*, in Proceedings of the International Conference on Learning Representations, 2020.
[69] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis, *Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators*, Nature Mach. Intell., 3 (2021), pp. 218–229.
[70] J.-M. Lueckmann, J. Boelts, D. Greenberg, P. Goncalves, and J. Macke, *Benchmarking simulation-based inference*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 343–351.
[71] A. L. Maas, A. Y. Hannun, and A. Y. Ng, *Rectifier nonlinearities improve neural network acoustic models*, in Proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language Processing, Vol. 28, 2013.
[72] A. Makkuva, A. Taghvaei, S. Oh, and J. Lee, *Optimal transport mapping via input convex neural networks*, in Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 6672–6681.
[73] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, *Least squares generative adversarial networks*, in Proceedings of the IEEE International Conference on Computer Vision, 2017.
[74] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini, *Sampling via Measure Transport: An Introduction*, in Handbook of Uncertainty Quantification, Springer, New York, 2016, pp. 1–41.
[75] R. J. McCann, *Existence and uniqueness of monotone measure-preserving maps*, Duke Math. J., 80 (1995), pp. 309–323.
[76] M. Mirza and S. Osindero, *Conditional Generative Adversarial Nets*, preprint, arXiv:1411.1784, 2014.
[77] S. Mo, C. Kim, S. Kim, M. Cho, and J. Shin, *Mining gold samples for conditional gans*, Adv. Neural Inf. Process. Syst., 32 (2019).

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

[78] A. Müller, *Integral probability metrics and their generating classes of functions*, Adv. Appl. Probab., 29 (1997), pp. 429–443.
[79] B. Muzellec and M. Cuturi, *Subspace detours: Building transport plans that are optimal on subspace projections*, Adv. Neural Inf. Process. Syst., 32 (2019).
[80] A. Y. Ng and M. I. Jordan, *On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes*, in Advances in Neural Information Processing Systems, 2002, pp. 841–848.
[81] S. Nowozin, B. Cseke, and R. Tomioka, *f-GAN: Training generative neural samplers using variational divergence minimization*, in Advances in Neural Information Processing Systems, Vol. 29, 2016.
[82] D. Onken, S. Wu Fung, X. Li, and L. Ruthotto, *OT-flow: Fast and accurate continuous normalizing flows via optimal transport*, in Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021.
[83] G. Papamakarios and I. Murray, *Fast ε-free inference of simulation models with Bayesian conditional density estimation*, in Advances in Neural Information Processing Systems, 2016.
[84] G. Papamakarios, E. T. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, *Normalizing flows for probabilistic modeling and inference*, J. Mach. Learn. Res., 22 (2021), pp. 1–64.
[85] G. Papamakarios, T. Pavlakou, and I. Murray, *Masked autoregressive flow for density estimation*, in Advances in Neural Information Processing Systems, Vol. 30, 2017.
[86] M. D. Parno and Y. M. Marzouk, *Transport map accelerated Markov chain Monte Carlo*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 645–682.
[87] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, *Context encoders: Feature learning by in-painting*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[88] S. T. Radev, U. K. Mertens, A. Voss, L. Ardizzone, and U. Köthe, *BayesFlow: Learning complex stochastic models with invertible neural networks*, IEEE Trans. Neural Netw. Learn. Syst., 33 (2022), pp. 1452–1466.
[89] D. Ray, H. Ramaswamy, D. V. Patel, and A. A. Oberai, *The efficacy and generalizability of conditional GANs for posterior inference in physics-based inverse problems*, Numer. Algebra Control Optim., 14 (2024), pp. 160–189.
[90] D. Rezende and S. Mohamed, *Variational inference with normalizing flows*, in Proceedings of the International Conference on Machine Learning, 2015.
[91] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 2015.
[92] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Springer, Berlin, 2005.
[93] M. Rosenblatt, *Remarks on a multivariate transformation*, Ann. Math. Statist., 23 (1952), pp. 470–472.
[94] J. Rothfuss, F. Ferreira, S. Walther, and M. Ulrich, *Conditional Density Estimation with Neural Networks: Best Practices and Benchmarks*, preprint, arXiv:1903.00954, 2019.
[95] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Springer, New York, 2015.
[96] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel, *Large scale optimal transport and mapping estimation*, in Proceedings of the International Conference on Learning Representations, 2018.
[97] M. Shiga, V. Tangkaratt, and M. Sugiyama, *Direct conditional probability density estimation with sparse feature selection*, Mach. Learn., 100 (2015), pp. 161–182.
[98] A. Siahkoohi, G. Rizzuti, M. Louboutin, P. A. Witte, and F. J. Herrmann, *Preconditioned training of normalizing flows for variational inference in inverse problems*, in Proceedings of the Third Symposium on Advances in Approximate Bayesian Inference, 2021.
[99] J. Song, A. Vahdat, M. Mardani, and J. Kautz, *Pseudoinverse-guided diffusion models for inverse problems*, in Proceedings of the International Conference on Learning Representations, 2023.
[100] J. Song, S. Zhao, and S. Ermon, *A-NICE-MC: Adversarial training for MCMC*, Adv. Neural Inf. Process. Syst., 30 (2017).
[101] Y. Song, L. Shen, L. Xing, and S. Ermon, *Solving inverse problems in medical imaging with score-based generative models*, in Proceedings of the 10th International Conference on Learning Representations, 2022.

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

[102] A. Spantini, R. Baptista, and Y. Marzouk, *Coupling techniques for nonlinear ensemble filtering*, SIAM Rev., 64 (2022), pp. 921–953.
[103] R. Strauss and J. B. Oliva, *Arbitrary conditional distributions with energy*, Adv. Neural Inf. Process. Syst., 34 (2021), pp. 752–763.
[104] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.
[105] A. Taghvaei and B. Hosseini, *An optimal transport formulation of bayes' law for nonlinear filtering algorithms*, in Proceedings of the 61st Conference on Decision and Control (CDC), IEEE, 2022, pp. 6608–6613.
[106] C. Villani, *Optimal Transport: Old and New*, Springer, New York, 2009.
[107] S. J. Vollmer, *Dimension-independent MCMC sampling for inverse problems with non-Gaussian priors*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 535–561.
[108] S. Wang and Y. Marzouk, *On Minimax Density Estimation via Measure Transport*, preprint, arXiv:2207.10231, 2022.
[109] J. Whang, E. Lindgren, and A. Dimakis, *Composing normalizing flows for inverse problems*, in Proceedings of the International Conference on Machine Learning, PMLR, 2021, pp. 11158–11169.
[110] J. Zech and Y. Marzouk, *Sparse approximation of triangular transports, part* I*: The finite-dimensional case*, Constr. Approx., 55 (2022), pp. 919–986.
[111] J. Zech and Y. Marzouk, *Sparse approximation of triangular transports, part* II*: The infinite-dimensional case*, Constr. Approx., 55 (2022), pp. 987–1036.
[112] E. Zeidler, *Nonlinear Functional Analysis and Its Applications:* II*/B:Nonlinear Monotone Operators*, Springer, New York, 2013.
[113] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, *Advances in variational inference*, IEEE Trans. Pattern Anal. Mach. Intell., 41 (2018), pp. 2008–2026.
[114] X. Zhou, Y. Jiao, J. Liu, and J. Huang, *A deep generative approach to conditional sampling*, J. Amer. Statist. Assoc., (2022), pp. 1–12.

Downloaded 09/03/24 to 205.175.118.107 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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