USING MMD GANS TO CORRECT PHYSICS MODELS AND IMPROVE BAYESIAN PARAMETER ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian parameter estimation methods are robust techniques for quantifying properties of physical systems which cannot be observed directly. In estimating such parameters, one first requires a physics model of the phenomenon to be studied. Often, such a model follows a series of assumptions to make parameter inference feasible. When simplified models are used for inference, however, systematic differences between model predictions and observed data may propagate throughout the parameter estimation process, biasing inference results. In this work, we use generative adversarial networks (GANs) based on the maximum mean discrepancy (MMD) to learn small stochastic corrections to physics models in order to minimize inference bias. We further propose a hybrid training procedure utilizing both the MMD and the standard GAN objective functionals. We demonstrate the ability to learn stochastic model corrections and eliminate inference bias on a toy problem wherein the true data distribution is known. Subsequently, we apply these methods to a mildly ill-posed inference problem in magnetic resonance imaging (MRI), showing improvement over an established inference method. Finally, because 3D MRI images often contain millions of voxels which would each require parameter inference, we train a conditional variational autoencoder (CVAE) network on the corrected MRI physics model to perform fast inference and make this approach practical.

1 INTRODUCTION

Bayesian parameter estimation methods are and robust techniques for quantifying properties of physical systems which cannot be observed directly (von Toussaint, 2011). In order to estimate such parameters, one first needs to develop a physics model of the phenomenon to be studied. This process requires deep domain-specific knowledge. For all but the most basic of physical systems, a series of simplifying assumptions on the physics of the system is required to make parameter inference computationally feasible. Examples of such methodologies include perturbation theory, in which higher order terms of Taylor series are dropped; mean-field theory, in which interactions involving many degrees of freedom are replaced by averaged approximations; and (stochastic) differential equations, when used as continuous limits of discrete stochastic processes. The price one pays for utilizing a given approximation is highly problem dependent. When approximate models are used for parameter inference, the mismatch between model predictions and data may propagate throughout the inference process and lead to bias and misestimations of the inferred parameters. In this work, we addressed the question **how can one improve a physics model for the purpose of parameter inference in a computationally inexpensive way?**

In order to make use of the physics contained in approximate but successful models, we propose a framework inspired by generative adversarial networks (GANs) to augment a model output with small stochastic corrections learned entirely from unlabeled observations. Here, GAN generators add stochastic corrections to the model outputs and discriminators critique the generator outputs. We investigate the use of GANs based on the optimized maximum mean discrepancy (MMD) for this task, as the MMD is a powerful discriminator between distributions (Bounliphone et al., 2016; Sutherland et al., 2019). In section 3, we describe the maximum likelihood estimation (MLE) procedure used for baseline parameter estimation, the MMD, traditional GANs, and MMD GANs. We further propose a hybrid training procedure which alternates between optimizing MMD GAN and standard GAN objective functionals. In section 4, we describe two experiments in order to compare

the ability of MMD GANs, traditional GANs, and hybrid-trained MMD GANs to improve parameter inference and to match the data distributions. The first experiment examines a toy problem wherein the true data distribution is known and the physics model to be corrected is deliberately misspecified. Subsequently, we consider a mildly ill-posed inference problem in magnetic resonance imaging (MRI), wherein inference results are compared with an established inference method.

Application to magnetic resonance imaging In MRI, the modelling of MR time signals is always challenged by impediments such as the complexity of the underlying biological tissue, MRI scanner hardware limitations, and spatiotemporal resolution limits. However, for (3+1)D MRI images with 1D MR time signals acquired for each voxel in a 3D volume, each time signal is usually well approximated by a superposition of relatively simple functions that are specific to tissue parameters. Nevertheless, such superpositions still bias the parameter inference required to produce an image, particularly for parameters which are sensitive to noise. In order to mitigate these issues, we investigate learning stochastic model corrections to aid inference. Additionally, parameter estimation is computationally expensive in MRI. In brain imaging, scan volumes are discretized into millions of voxels on the order of cubic millimetres, resulting in millions of nonnegative least squares (NNLS) or maximum likelihood estimation (MLE) inference problems. Therefore, fast approximate data-driven inference methods are appealing. Recently, conditional variational autoencoders (CVAEs) were used to accelerate posterior sampling of gravitational wave event source parameters by 6 orders of magnitude (Gabbard et al., 2019). Here, we show that CVAE inference trained on MMD GAN corrected toy model signals generalizes to signals drawn from the true distribution. Then, we perform a similar experiment using an MMD GAN as a source of realistic MRI signals, comparing CVAE inference results to an established NNLS-based inference method (Prasloski et al., 2012; Doucette et al., 2020) as well as a more expensive MLE method.

This work lies between two classes of GAN applications: GANs which are black-box functions mapping random noise to data samples, providing little understanding of the landscape of the input space, and GANs with well-understood input spaces, e.g. mapping images to images. By controlling the size of learned stochastic corrections to physics models, we explicitly trade-off between matching the goal data distribution and retaining model interpretability: smaller corrections give better understood input spaces, while larger corrections allow for better matching of the data distribution.

2 Related work

Applications of GANs in physics has been of increasing interest recently. For use as black-box samplers, de Oliveira et al. (2017) trained GANs for the generation of physically realistic 2D radiation patterns of high energy particle collisions from simulated data. Unlike natural images, here the target distribution is highly sparse, non-smooth, takes values ranging over several orders of magnitude, and much of the interesting physics occurs in the tail of the distributions. Nevertheless, the GAN learned to reproduce the data distribution to high accuracy, as well as the distributions of several low dimensional physics-inspired quantities derived from the images. In follow up work, Paganini et al. (2018) used GANs to emulate the distribution of simulated 3D particle showers in multi-layer calorimeters at CERN's large hadron collider. These studies saw speedups in sampling of up to 5 orders of magnitude compared to traditional simulations. GANs have also been used for inference problems in imaging which have well-understood input spaces arising from physics. In MRI image reconstruction, GANs were used to map undersampled image data – acquired in the Fourier domain – directly to reconstructed images, obtaining higher accuracy than traditional methods (Yang et al., 2018; Quan et al., 2018). Similarly, Hammernik et al. (2018) proposed using GANs with generators imbued with physics information for MRI reconstruction. There, the generator is a variational network (Kobler et al., 2017) – a deep network which learns to invert a variational problem – where the variational objective encodes the MRI physics.

3 Methods

Maximum likelihood estimation with Rician noise As a baseline method for parameter inference, we implement maximum likelihood estimation (MLE) for inferring parameters $\theta \in \mathbb{R}^{N_\theta}$ from noisy observations $Y \in \mathbb{R}^n$. As is the case for MRI signal measurements, we take Y to be described by independent Rician (Rice, 1944) distributions $Y_i \sim \mathrm{Rice}(X_i, \epsilon_i)$ conditional on $X = F(\theta)$,

where $F : \mathbb{R}^{N_\theta} \to \mathbb{R}^n$ is a mathematical model describing the underlying physics. We are interested in two cases for ϵ: $\epsilon_i \equiv \epsilon_0$ unknown and to be determined from the MLE procedure, or ϵ_i fixed and possibly X-dependent. In either case, θ and possibly ϵ are determined by solving

$$\underset{\theta,\,\epsilon}{\arg\min} \; -\sum_{j=1}^{n} \log P(Y_j \,|\, X_j, \epsilon_j) \quad \text{where} \quad X = F(\theta) \tag{1}$$

$P(y \,|\, \nu, \epsilon) = \frac{y}{\epsilon^2} \exp\left(\frac{-(y^2+\nu^2)}{2\epsilon^2}\right) I_0\left(\frac{y\nu}{\epsilon^2}\right)$ is the likelihood of y under $\mathrm{Rice}(\nu, \epsilon)$, where I_0 is the modified Bessel function of the first kind with order zero.

Generative adversarial networks and the maximum mean discrepancy GANs (Goodfellow et al., 2014) consist of two players, a generator and a discriminator, formulating unsupervised learning tasks as contests in which the discriminator criticizes the generator. We consider GANs for the task of improving an initial distribution $\mathbb{P}_X : \mathcal{X} \to \mathbb{R}_+$ toward a goal distribution $\mathbb{P}_Y : \mathcal{X} \to \mathbb{R}_+$ when only samples $X \sim \mathbb{P}_X$ and $Y \sim \mathbb{P}_Y$ are available. Here, X are simulated data, Y are observed data, and $\mathcal{X}$ is the data domain. Let $G : \mathcal{X} \to (\mathcal{X} \to \mathbb{R}_+)$ be a generator mapping $X \sim \mathbb{P}_X$ onto distributions $\mathbb{P}_{\hat{Y}} : \mathcal{X} \to \mathbb{R}_+$. In this formulation, $\hat{Y} \sim G(X)$ are samples from a learned distribution conditioned on X, with $\hat{Y}$ representing e.g. realizations of noisy data Y. We then consider two forms of GANs. First, the traditional form with $D : \mathcal{X} \to (0, 1)$ where G and D are trained via alternating gradient descent and ascent steps on the joint objective (Goodfellow et al., 2014)

$$\begin{cases} \min\limits_{G} & \mathbb{E}_{X\sim\mathbb{P}_X, \hat{Y}\sim G(X)} & \log(1 - D(\hat{Y})) \\ \max\limits_{D} & \mathbb{E}_{X\sim\mathbb{P}_X, Y\sim\mathbb{P}_Y, \hat{Y}\sim G(X)} & \log(D(Y)) + \log(1 - D(\hat{Y})). \end{cases} \tag{2}$$

Second, GANs based on the integral probability metric $\mathrm{MMD}(\mathbb{P}_X, \mathbb{P}_Y)$, defined in terms of a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Given batches of m samples $\mathbf{X} \sim \mathbb{P}_X^m$ and $\mathbf{Y} \sim \mathbb{P}_Y^m$,

$$\widehat{\mathrm{MMD}}_U^2(\mathbf{X}, \mathbf{Y}) \coloneqq \frac{1}{m(m-1)} \sum_{i\neq j}^{m} \left[k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(X_j, Y_i)\right] \tag{3}$$

is an unbiased estimator of $\mathrm{MMD}^2(\mathbb{P}_X, \mathbb{P}_Y)$ with nearly minimal variance among unbiased estimators (Gretton et al., 2012), and is a U statistic; a similar expression for the variance $\hat{V}_m(\mathbf{X}, \mathbf{Y}) \coloneqq \mathrm{Var}\,\widehat{\mathrm{MMD}}_U^2(\mathbf{X}, \mathbf{Y})$ due to Sutherland (2019) is given in appendix A. The MMD is used in both generator and discriminator contexts (Dziugaite et al., 2015; Li et al., 2015): the generator is trained to minimize the MMD, and the discriminator, i.e. the reproducing kernel k, is trained to maximize the MMD. Or alternatively, Sutherland et al. (2019) show that maximizing the t-statistic estimator

$$\hat{t}(\mathbf{X}, \mathbf{Y}) = \widehat{\mathrm{MMD}}_U^2(\mathbf{X}, \mathbf{Y}) \Big/ \sqrt{\hat{V}_m(\mathbf{X}, \mathbf{Y})} \tag{4}$$

asymptotically maximizes the power of rejecting the null hypothesis $H_0 : \mathbb{P}_X = \mathbb{P}_Y$ of a permutation test on data $\mathbf{X}$ and $\mathbf{Y}$ using the test statistic $m \cdot \widehat{\mathrm{MMD}}_U^2$. Therefore, $\hat{t}$ provides a differentiable proxy for permutation test power which the discriminator can maximize. Letting $\hat{\mathbf{Y}} \sim \Pi_{i=1}^m G(\mathbf{X}_i)$, the joint objective for MMD GANs analogous to equation (2) is (Sutherland et al., 2019):

$$\begin{cases} \min\limits_{G} & \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_X^m, \mathbf{Y}\sim\mathbb{P}_Y^m, \hat{\mathbf{Y}}\sim\Pi_{i=1}^m G(\mathbf{X}_i)} & \widehat{\mathrm{MMD}}_U^2(\hat{\mathbf{Y}}, \mathbf{Y}) \\ \max\limits_{k} & \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_X^m, \mathbf{Y}\sim\mathbb{P}_Y^m, \hat{\mathbf{Y}}\sim\Pi_{i=1}^m G(\mathbf{X}_i)} & \widehat{\mathrm{MMD}}_U^2(\hat{\mathbf{Y}}, \mathbf{Y}) \quad \textit{or} \quad \hat{t}(\hat{\mathbf{Y}}, \mathbf{Y}). \end{cases} \tag{5}$$

Model architectures In this work, $\mathcal{X} = \mathbb{R}^n$ and the reproducing kernel k is the mean of N_{bw} Gaussian basis functions with unique bandwidths for each coordinate of the inputs $X, Y \in \mathbb{R}^n$,

$$k(X, Y) = \frac{1}{N_{bw}} \sum_{i=1}^{N_{bw}} \exp\left(-\sum_{j=1}^{n} \frac{(X_j - Y_j)^2}{2\sigma_{i,j}^2}\right), \tag{6}$$

where $\sigma \in \mathbb{R}_+^{N_{bw}\times n}$ is the matrix of kernel bandwidths. These bandwidths are the free parameters of the kernel k which are optimized during the maximization over k in equation (5).

Algorithm 1: Training procedure for hybrid MMD GANs. G is the generator, D is the discriminator, and k is the reproducing kernel which defines the MMD.

for N_{epochs} *training epochs* **do**
  every k_{rate} *epochs* **do**
    Draw m samples of $\hat{Y}$ and Y and train k according to equation (5)
  for $N_{batches}$ *minibatches* **do**
    Draw m samples of $\hat{Y}$ and Y and train G according to equation (5)
    every GAN_{rate} *epochs* **do**
      for D_{steps} *steps* **do**
        Draw m samples of $\hat{Y}$ and Y and train D according to equation (2)
      Draw m samples of $\hat{Y}$ and train G according to equation (2)

The discriminator D in equation (2) is taken to be a fully connected neural network with N_d hidden layers, N_h hidden nodes per layer, ReLU activation on the hidden layers, and sigmoid activation on the final scalar output. The generator G produces samples $\hat{Y} \sim G(X)$ such that the components $\hat{Y}_i$ follow Rician distributions. G is defined implicitly by the sampling scheme

$$\begin{cases} \hat{Y}_i \sim \mathrm{Rice}(\nu_i, \epsilon_i) \\ \nu = |X + g_\delta(X)| \\ \log \epsilon = g_\epsilon(X) \end{cases} \tag{7}$$

where $\nu, \epsilon \in \mathbb{R}^n$, $\log$ and $|\cdot|$ are applied elementwise, and the functions $g_\delta : \mathbb{R}^n \to (-\delta, \delta)^n$ and $g_\epsilon : \mathbb{R}^n \to (\log \epsilon^-, \log \epsilon^+)^n$ are parameterized neural networks. The architectures of g_δ and g_ϵ are identical to that of the discriminator D except for their final layers, which have n output nodes and use $\tanh$ activations functions linearly scaled to the ranges $(-\delta, \delta)$ and $(\log \epsilon^-, \log \epsilon^+)$, respectively.

This parameterization of G permits only small deterministic corrections $g_\delta(X)$ for each X which are bounded *a priori* by a fixed constant δ. That is, one explicitly places limits on the size of the non-physical learned corrections. The noise level ϵ is additionally allowed to be data-dependent, as in MRI applications it is common for both signal-to-noise ratio and ϵ to vary with signal amplitude and across scan volumes. The range of g_ϵ is also bound by the noise level limits ϵ^- and ϵ^+.

Hybrid training procedure We propose a hybrid training procedure which makes use of both traditional GAN and MMD GAN objective functionals. Let a hybrid GAN consist of a generator G, discriminator D, and reproducing kernel k. Then, updates to k and G via equation (5) are interleaved with updates to D and G via equation (2) as described in algorithm 1. The k update is pulled out of the minibatch loop and applied only every k_{rate} epochs to limit the ability of k to overwhelm G and cause its gradients to vanish; regularizing k is a difficult problem for which we took the simplest approach, though many sophisticated techniques exist (Arbel et al., 2018; Li et al., 2015; Bińkowski et al., 2018). The traditional GAN updates to G and D are applied every GAN_{rate} epochs, with D updated D_{steps} times for every G update.

Conditional variational autoencoders In this work, CVAEs are used to perform fast approximate sampling of the Bayesian posterior function $P(\theta \mid \hat{Y})$ where $\hat{Y} \sim G(X)$, G is a pre-trained generator, and X is output from an uncorrected physics model which depends on θ. The CVAE architecture we use is exactly that of Gabbard et al. (2019); see appendix B.2 for details.

4 Experiments

Toy physics model We first study a toy physical system with a known data distribution. Consider

$$F_\beta(t; \theta) = \left(a_0 + a_1 \sin^\beta(2\pi f t - \phi)\right) \exp(-t/\tau) \tag{8}$$

where $F_\beta : \mathbb{R} \to \mathbb{R}$ is an exponential decay curve modulated by a periodically varying amplitude parameterized by $\theta = [f, \phi, a_0, a_1, \tau] \in \mathbb{R}^5$. The observations are time signals $Y \in \mathbb{R}^n$ where

$n = 128$, $Y_i \sim \text{Rice}(\nu_i, \epsilon_0)$, $\nu_i = F_{\beta=2}(t_i; \theta)$, $t_i = i-1$, and $\epsilon_0 = 0.01$. The behaviour of this process, however, is modelled by $X \in \mathbb{R}^n$ where $X_i = F_{\beta=4}(t; \theta)$, purposefully introducing a systematic bias into the recovered parameters $\hat{\theta}$. We first infer $\hat{\theta}$ and $\hat{\epsilon}_0$ via equation (1) using both the true $\beta = 2$ and misspecified $\beta = 4$ models to establish best and worst case inference performance. Then, generators defined by equation (7) are trained to learn corrected signals $\hat{Y}$ via equation (2), equation (5), and algorithm 1, with the trained generators used to infer $\hat{\theta}$ via equation (1) for comparison. Lastly, we show that a CVAE model trained on $\hat{Y} \sim G(X)$ generalizes to the true data Y. In all experiments, the prior space $\mathbb{P}_\theta$ is defined by the priors $f \sim U(\frac{1}{64}, \frac{1}{32})$, $\phi \sim U(0, \frac{\pi}{2})$, $a_0 \sim U(\frac{1}{4}, \frac{1}{2})$, $a_1 \sim U(\frac{1}{10}, \frac{1}{4})$, and $\tau \sim U(16, 128)$; $U(a, b)$ is the uniform distribution on (a, b). Training data sets for X and Y were created from two draws of 102 400 unique $\theta \sim \mathbb{P}_\theta$ samples. Separate validation and testing data sets were similarly created with 10 240 samples each.

MRI physics model Here, we study an MRI physics model in which multi spin-echo MRI time signals are modelled as the superposition of two component signals. Let $Y \in \mathbb{R}^n$ be an MRI time signal consisting of n measurements at uniformly spaced sample times t_i, and let $F : \mathbb{R} \to \mathbb{R}$ be

$$F(t; \theta) = \sum_{\ell=1}^{2} A_\ell \ \text{EPG}(t, \alpha, T_{2,\ell}) \tag{9}$$

where $\text{EPG}(t, \alpha, T_{2,\ell})$ is a component signal computed using the extended phase graph (EPG) algorithm (Hennig, 1988). $\text{EPG}(t, \alpha, T_{2,\ell})$ is roughly equal to $\exp(-t/T_{2,\ell})$, with MRI physics corrections due to the spin flip angle α. We adopt the convention $T_{2,1} \leq T_{2,2}$, denoting $T_{2,short} = T_{2,1}$, $T_{2,long} = T_{2,2}$, $A_{short} = A_1$, and $A_{long} = A_2$. These MRI signal parameters are of considerable academic and clinical interest, for example in myelin water imaging (Mackay et al., 1994; Whittall & MacKay, 1989) for brain research (Wright et al., 2016; Weber et al., 2020), and in luminal water imaging for prostate cancer research (Sabouri et al., 2017). The measured signal is modelled as $Y_i \sim \text{Rice}(X_i, \epsilon_0)$ where $X \in \mathbb{R}^n$, $X_i = F(t_i; \theta)$, and $\theta = [\alpha, T_{2,short}, T_{2,long}, A_{short}, A_{long}] \in \mathbb{R}^5$, with $\hat{\theta}$ and $\hat{\epsilon}_0$ inferred using equation (1). Then, as in the toy model experiments, we train generators defined by equation (7) via equation (2), equation (5), and algorithm 1, producing corrected MRI models. MLE is performed to infer $\hat{\theta}$ using these corrected models, and a CVAE model is trained using signals drawn from the MMD GAN generator. In all experiments, the prior space $\mathbb{P}_\theta$ consists of the $\hat{\theta}$ values resulting from the MLE fits of $F(t; \theta)$ to the data Y. During fitting, $\hat{\theta}$ was constrained as: $\alpha \in [50°, 180°]$, $T_{2,short} \in [8\,\text{ms}, 1000\,\text{ms}]$, $T_{2,long} - T_{2,short} \in [8\,\text{ms}, 1000\,\text{ms}]$, and $A_{short}, A_{long} \in [0, \infty)$. Half of these $\hat{\theta}$ and corresponding X and Y values are used for training data, with the remaining half split evenly between validation and testing data. The MRI data used for this experiment is from a brain scan using a CPMG sequence (Whittall et al., 1997) on a 3 T MR system (Ingenia Elition, Philips Medical Systems, Best, The Netherlands) from a healthy volunteer giving written and informed consent, approved by our university ethics board. The extracted brain volume consists of 821 145 signals $Y \in \mathbb{R}^n$ acquired with $n = 48$ samples at times $t_i = i \cdot \text{TE}$, echo spacing $\text{TE} = 8\,\text{ms}$, and spatial resolution of $0.96 \times 0.96 \times 2.5\text{mm}^3$.

This two-component model is a simplification of an established multicomponent model (Prasloski et al., 2012) wherein inference is performed using a regularized NNLS-based technique with a fixed spectrum of many – typically 40 or more – $T_{2,\ell}$ components. The resulting amplitudes A_ℓ together with $T_{2,\ell}$ are collectively referred to as the T_2 *distribution* of the MRI signal, and can be thought of as a regularized inverse Laplace transform with EPG basis functions substituting for exponential functions. This method, herein referred to as NNLS, is used for further comparison of inference results. Note that while this method is well established, it is not optimal; for instance, it implicitly assumes a uniform Gaussian noise distribution as opposed to a Rician distribution, which is only justified for high signal-to-noise ratios.

Training All models were trained for two days using the Adam (Kingma & Ba, 2014) optimizer with default momentum. Hyperparameter sweeps were performed to optimize the following: learning rate η, k and GAN training rates k_{rate} and GAN_{rate}, number of D updates per G update D_{steps}, number of σ bandwidths N_{bw}, kernel loss $\widehat{\text{MMD}}_U^2$ or $\hat{t}$, number of hidden layers N_h and hidden nodes N_d, and batch size m. For toy (MRI) models, $\delta = 0.1\ (0.025)$, $\log \epsilon \in [-8, -2]\ ([-6, -3])$, $\log \sigma \in [-8, 4]\ ([-8, 4])$. GAN performance was evaluated on maximum log likelihood value on validation data. In depth training details for these and CVAE models is given in appendix B.2.

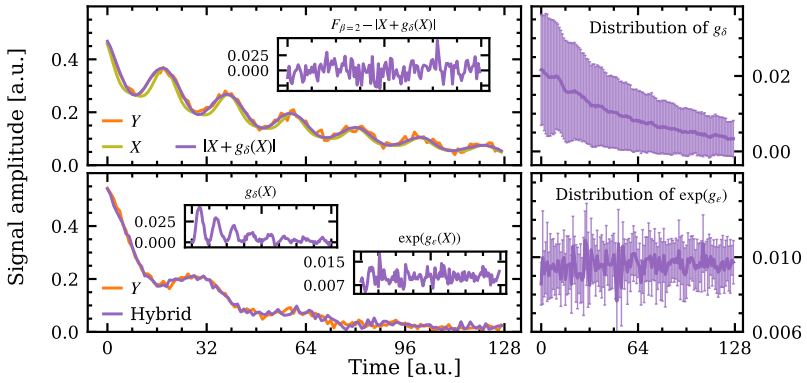

Figure 1: Example toy model signals and corresponding corrections learned via the hybrid training procedure in algorithm 1. Top-left: noisy true signal Y overlaid with noiseless uncorrected X and corrected $|X + g_\delta(X)|$ signals, with X given by $F_{\beta=4}$ in equation (8); difference between true $F_{\beta=2}$ and learned $|X + g_\delta(X)|$ is inset. Bottom-left: noisy true signal Y overlaid with noisy corrected signal (Hybrid); learned correction $g_\delta(X)$ and noise level $\exp(g_\epsilon(X))$ is inset. Top-right & bottom-right: distributions of learned g_δ and $\exp(g_\epsilon)$ over the validation X data.

Table 1: Comparison of inference results for the toy physics problem. Mean absolute errors for inferred parameters are shown as percentages relative to respective uniform prior distribution widths.

	$F_{\beta=4}$	GAN	MMD GAN	Hybrid GAN	CVAE	$F_{\beta=2}$
Log likelihood	371.7 ± 1.1	409.8 ± 0.4	409.6 ± 0.4	408.4 ± 0.4	–	413.5 ± 0.3
RMSE/ϵ_0	0.790 ± 0.008	0.239 ± 0.002	0.257 ± 0.002	0.250 ± 0.002	–	0.189 ± 0.002
$f = \theta_1$	1.13 ± 0.05	0.79 ± 0.04	0.79 ± 0.03	0.81 ± 0.04	0.77 ± 0.04	0.69 ± 0.03
$\phi = \theta_2$	2.45 ± 0.07	1.62 ± 0.05	1.53 ± 0.04	1.56 ± 0.05	1.44 ± 0.04	1.32 ± 0.04
$a_0 = \theta_3$	8.72 ± 0.10	1.62 ± 0.04	1.32 ± 0.03	1.48 ± 0.04	1.17 ± 0.03	1.08 ± 0.03
$a_1 = \theta_4$	5.45 ± 0.12	3.20 ± 0.10	3.55 ± 0.10	3.41 ± 0.10	3.16 ± 0.10	3.00 ± 0.09
$\tau = \theta_5$	0.76 ± 0.03	0.75 ± 0.02	0.65 ± 0.02	0.78 ± 0.03	0.57 ± 0.02	0.54 ± 0.02

5 Results

Toy physics model Figure 1 shows example toy model signals before and after adding learned corrections, and table 1 compares toy problem inference results using six models: MLE with the uncorrected $F_{\beta=4}$ model, estimating $\hat{\epsilon}_0$; MLE with GAN, MMD GAN, and hybrid GAN generator models, with learned $\hat{\epsilon}$; posterior sampling with a CVAE trained on MMD GAN outputs; and MLE with the true $F_{\beta=2}$ model, using the true $\epsilon_0 = 0.01$. 1000 signals randomly chosen from the validation data are used for model comparison. Log likelihood values are optimized via equation (1). Root-mean-square error (RMSE) values relative to ϵ_0 are computed between the model fits without added noise, i.e. $|X + g_\delta(X)|$, and the true model $F_{\beta=2}$ without noise. Mean absolute error is shown for each inferred parameter relative to the width of the corresponding uniform prior distribution. Note that CVAE-inferred parameters are nearly as accurate as the perfectly specified model $F_{\beta=2}$.

MRI physics model Figure 2 shows learned generator outputs for the MRI physics model, signal amplitude histograms using bin widths determined by the MRI digitizer and populated with the $48 \cdot 205\,287$ validation data samples, and the mean over validation signal T_2 distributions computed using the NNLS inference method. Table 2 compares inference results in an identical manner as in table 1, using F from equation (9) as the uncorrected model, but without a ground truth. In lieu of a true model, we compute the ℓ_1, ℓ_2, χ^2, and Wasserstein distances between model and Y histograms (Pele & Werman, 2010), as well as the ℓ^2 norm of T_2 distribution differences. Lastly, as visualized in figure 3, inferred parameters are compared with NNLS. The spin flip angle α is compared directly. The geometric means of the T_2 distribution from NNLS and the $T_{2,\ell}$ values from F are compared, both weighted by the corresponding amplitudes $A_{2,\ell}$. Myelin water fraction (MWF) (Mackay et al., 1994) and $T_{2,long}$ are also visualized in figure 3. For NNLS, MWF is the fraction of A_ℓ contained in $T_{2,\ell} \leq 40\,\text{ms}$, and $T_{2,long}$ is the geometric mean of $T_{2,\ell} \geq 40\,\text{ms}$; for CVAE, MWF is $A_{short}/(A_{short} + A_{long})$ weighted by the logistic function $\sigma((30\,\text{ms} - T_{2,short})/10\,\text{ms})$.

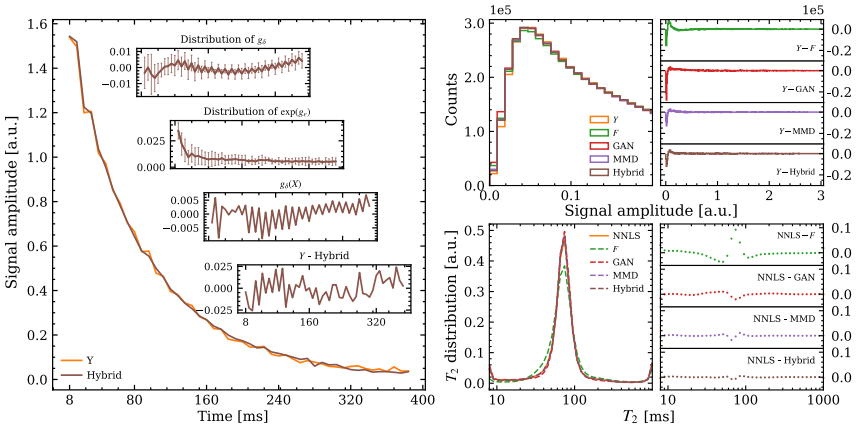

Figure 2: Left: example noisy MRI signal Y compared with noisy corrected signal (Hybrid), trained via the hybrid training algorithm 1; distributions of learned g_δ and $\exp(g_\epsilon)$ over the validation X data is inset, as well as the learned correction $g_\delta(X)$ and signal difference for the data shown. Top-right: comparison of true Y vs. learned signal amplitude distributions for the uncorrected $F(t;\theta)$ from equation (9), GAN, MMD GAN, and hybrid-trained GAN signal models over the validation X and Y data, as well as the distribution differences compared to Y. Bottom-right: comparison of true vs. learned T_2 distributions – analagous to regularized inverse Laplace transforms – and T_2 distribution differences compared to Y for the same four signal models. The hybrid training algorithm produces signal amplitude and T_2 distributions which are most similar to the true distributions.

Table 2: Comparison of inference results and distribution similarity for the MRI physics problem. Maximum likelihood estimation is performed for the uncorrected $F(t;\theta)$, GAN, MMD GAN, and hybrid-trained GAN signal models, with the resulting log-likelihood shown. Spin flip angle α and geometric mean T_2 and compared with NNLS for each signal model, as well as CVAE; note that NNLS, while commonly used, is not a gold standard and small differences are expected. Five distribution similarity metrics are shown comparing the four signal models with the true data distributions; the hybrid-trained model performs best in all metrics.

	F	GAN	MMD GAN	Hybrid GAN	CVAE
Log likelihood	163.7 ± 0.6	174.0 ± 0.6	174.0 ± 0.5	175.4 ± 0.5	–
Spin flip angle α	$0.86 \pm 0.04°$	$4.87 \pm 0.21°$	$3.08 \pm 0.11°$	$3.59 \pm 0.16°$	$2.99 \pm 0.13°$
Geo. mean T_2	4.30 ± 0.26 ms	6.19 ± 0.35 ms	5.08 ± 0.29 ms	6.00 ± 0.32 ms	10.0 ± 0.5 ms
ℓ_1	143 655.0	147 934.0	94 282.0	90 786.0	–
ℓ_2	28 298.3	37 807.0	19 491.3	18 573.0	–
χ^2	2944.6	5339.2	1279.8	919.6	–
Wasserstein	390.1	342.2	253.5	243.3	–
T_2 distribution ℓ_2	0.141	0.038	0.024	0.014	–

6 Discussion

In this work, we trained three classes of generative adversarial networks to learn stochastic corrections to physics models and improve Bayesian parameter estimation. As seen in figure 1 and table 1, all three GAN models as well as the CVAE model trained on MMD GAN outputs result in near optimal inference performance for the toy problem, which has the advantage of having corrections that are exactly described by the same parameters θ as the uncorrected model. The MRI physics model does not have this luxury, and furthermore g_δ and $\exp(g_\epsilon)$ are on average only 1-2 percent as large as the signal intensity, yet figure 2 and table 2 show considerable improvement in inference and distribution similarity. Additionally, figure 3 shows that the robustly estimated parameters α and mean T_2 are consistent between the NNLS method and a CVAE trained on MMD GAN generator signals. The MWF, a biomarker for brain myelin content, depends strongly on fast decaying signal components making it sensitive to noise. The CVAE generates a more spatially uniform MWF map, which is expected from myelin biology, and resolves smaller structures compared to NNLS. Additionally, $T_{2,long}$ is misestimated by NNLS in the cerebrospinal fluid – the bright regions in the CVAE image – where it is known that $T_{2,long} \sim 1$ s.

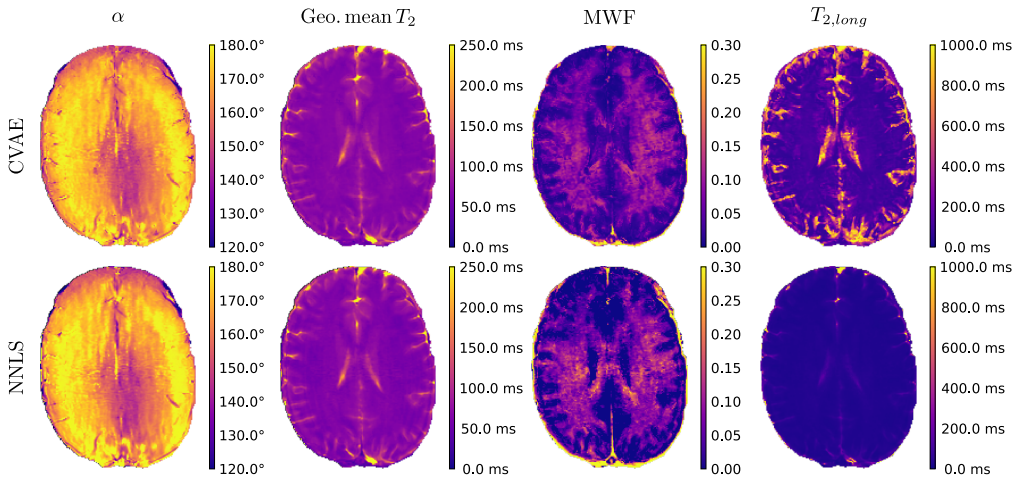

Figure 3: Example parameter maps computed using CVAE (top row) and NNLS (bottom row). Estimation of the spin flip angle α and geometric mean T_2 is relatively insensitive to noise, resulting in little difference between the two methods. MWF is harder to estimate due to its dependence on fast decaying signal components with inherently lower signal-to-noise ratio; CVAE produces more spatially uniform MWF compared to NNLS, which is expected from neurobiology. $T_{2,long}$ is misestimated by NNLS: slowly decaying signal components cannot be distinguished from noise in NNLS, resulting in underestimation of the long T_2 of the cerebrospinal fluid which is known to be $\sim 1\,\mathrm{s}$. The CVAE, trained using the correct Rician noise model, does not have this issue.

While training the different models, we observed MMD GANs to have less severe failure modes than traditional GANs. MMD GANs first learn the mean correction toward the target distribution, slowly recovering finer details thereafter; in the toy problem, smooth exponentially decaying corrections were learned first, with the oscillatory corrections following. Traditional GANs, however, would proceed more erratically toward the true distribution. On the other hand, noise amplitude distributions learned by traditional GANs appeared more physically plausible, exhibiting less spurious variations in time. In all, the ability of traditional GANs to critique individual signals with high fidelity and MMD GANs to discriminate between distributions motivated the hybrid training procedure in algorithm 1. For example, the distribution of $\exp(g_\epsilon)$ in figure 1 is much more homogeneous in time compared to MMD GANs; see figures 4 and 5 in appendix B. Note however that traditional GANs on their own are limited for this application: generators $G = G(X(\theta))$ depend only on θ and are unable to capture behaviour described by unmodelled parameters. For traditional GANs, discriminators eventually learn to exploit this mismatch, resulting in undesired behaviour such as the decreased distribution similarity observed in table 2. In the hybrid training routine, however, the discriminator acts as an adversarial regularizer to the MMD GAN objective, improving results.

Limitations and future work The experiments described in section 4 place explicit limits on the range of g_δ and g_ϵ, allowing one to control the size of the stochastic corrections learned by G. However, we have not shown formally that these corrections cannot substantially change the physical interpretation of inferred $\hat{\theta}$. For example, the inversion of the MRI model becomes ill-posed as $T_{2,short} \to T_{2,long}$ in equation (9), leading to inference which is sensitive to small changes in the MRI signal, such as those introduced by the learned corrections. In this case one could argue that such $\hat{\theta}$ had limited meaning in the first place, but in general there may be subtle failure modes.

For future work, one may consider allowing G to depend on additional nuisance parameters z. This would allow G to learn corrections driven by physics present in the goal distribution that is not described by θ, but could be characterized by z. During inference, inferred $\hat{\theta}$ would be forced to be independent of z; for example, a CVAE could be trained to pivot (Louppe et al., 2017) on z.

Lastly, while we have focused on physics motivated signal processing, this framework is fully general and may find applications to any system which well approximates a goal distribution but could benefit from learning controlled stochastic corrections from unlabeled data.

REFERENCES

Michael Arbel, Dougal Sutherland, Mikołaj Bińkowski, and Arthur Gretton. On gradient regularizers for MMD GANs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6700–6710. Curran Associates, Inc., 2018.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv:1801.01401 [cs, stat]*, March 2018.

Wacha Bounliphone, Eugene Belilovsky, Matthew B. Blaschko, Ioannis Antonoglou, and Arthur Gretton. A Test of Relative Similarity For Model Selection in Generative Models. *arXiv:1511.04581 [cs, stat]*, February 2016.

Luke de Oliveira, Michela Paganini, and Benjamin Nachman. Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis. *Computing and Software for Big Science*, 1(1):4, September 2017. ISSN 2510-2044. doi: 10.1007/s41781-017-0004-6.

Jonathan Doucette, Christian Kames, and Alexander Rauscher. DECAES - DEcomposition and Component Analysis of Exponential Signals. *Zeitschrift Fur Medizinische Physik*, May 2020. ISSN 1876-4436. doi: 10.1016/j.zemedi.2020.04.001.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via Maximum Mean Discrepancy optimization. *arXiv:1505.03906 [cs, stat]*, May 2015.

Hunter Gabbard, Chris Messenger, Ik Siong Heng, Francesco Tonolini, and Roderick Murray-Smith. Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy. *arXiv:1909.06296 [astro-ph, physics:gr-qc]*, September 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. ISSN 1533-7928.

Kerstin Hammernik, Erich Kobler, Thomas Pock, Michael P. Recht, Daniel K. Sodickson, and Florian Knoll. Variational Adversarial Networks for Accelerated MR Image Reconstruction. In *Proceedings of the 26th Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM), Paris, France*, June 2018.

J Hennig. Multiecho imaging sequences with low refocusing flip angles. *Journal of Magnetic Resonance (1969)*, 78(3):397–407, July 1988. ISSN 0022-2364. doi: 10.1016/0022-2364(88)90128-X.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December 2014.

Erich Kobler, Teresa Klatzer, Kerstin Hammernik, and Thomas Pock. Variational Networks: Connecting Variational Methods and Deep Learning. In *Pattern Recognition*, Lecture Notes in Computer Science, pp. 281–293. Springer, Cham, September 2017. ISBN 978-3-319-66708-9 978-3-319-66709-6. doi: 10.1007/978-3-319-66709-6_23.

Yujia Li, Kevin Swersky, and Richard Zemel. Generative Moment Matching Networks. *arXiv:1502.02761 [cs, stat]*, February 2015.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to Pivot with Adversarial Networks. *arXiv:1611.01046 [physics, stat]*, June 2017.

Alex Mackay, Kenneth Whittall, Julian Adler, David Li, Donald Paty, and Douglas Graeb. In vivo visualization of myelin water in brain by magnetic resonance. *Magnetic Resonance in Medicine*, 31(6):673–677, 1994. ISSN 1522-2594. doi: 10.1002/mrm.1910310614.

Michela Paganini, Luke de Oliveira, and Benjamin Nachman. Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multi-Layer Calorimeters. *Physical Review Letters*, 120(4):042003, January 2018. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.120.042003.

Ofir Pele and Michael Werman. The Quadratic-Chi Histogram Distance Family. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios (eds.), *Computer Vision – ECCV 2010*, Lecture Notes in Computer Science, pp. 749–762, Berlin, Heidelberg, 2010. Springer. ISBN 978-3-642-15552-9. doi: 10.1007/978-3-642-15552-9_54.

Thomas Prasloski, Burkhard Mädler, Qing-San Xiang, Alex MacKay, and Craig Jones. Applications of stimulated echo correction to multicomponent T2 analysis. *Magnetic Resonance in Medicine*, 67(6):1803–1814, 2012. ISSN 1522-2594. doi: 10.1002/mrm.23157.

Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. Compressed Sensing MRI Reconstruction Using a Generative Adversarial Network With a Cyclic Loss. *IEEE Transactions on Medical Imaging*, 37(6):1488–1497, June 2018. ISSN 1558-254X. doi: 10.1109/TMI.2018.2820120.

S. O. Rice. Mathematical Analysis of Random Noise. *Bell System Technical Journal*, 23(3):282–332, 1944. ISSN 1538-7305. doi: 10.1002/j.1538-7305.1944.tb00874.x.

Shirin Sabouri, Silvia D. Chang, Richard Savdie, Jing Zhang, Edward C. Jones, S. Larry Goldenberg, Peter C. Black, and Piotr Kozlowski. Luminal Water Imaging: A New MR Imaging T2 Mapping Technique for Prostate Cancer Diagnosis. *Radiology*, 284(2):451–459, April 2017. ISSN 0033-8419. doi: 10.1148/radiol.2017161687.

Dougal J. Sutherland. Unbiased estimators for the variance of MMD estimators. *arXiv:1906.02104 [cs, stat]*, June 2019.

Dougal J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy. *arXiv:1611.04488 [cs, stat]*, June 2019.

Udo von Toussaint. Bayesian inference in physics. *Reviews of Modern Physics*, 83(3):943–999, September 2011. doi: 10.1103/RevModPhys.83.943.

Alexander Mark Weber, Yuting Zhang, Christian Kames, and Alexander Rauscher. Myelin water imaging and R2* mapping in neonates: Investigating R2* dependence on myelin and fibre orientation in whole brain white matter. *NMR in biomedicine*, 33(3):e4222, March 2020. ISSN 1099-1492. doi: 10.1002/nbm.4222.

Kenneth P Whittall and Alexander L MacKay. Quantitative interpretation of NMR relaxation data. *Journal of Magnetic Resonance (1969)*, 84(1):134–152, August 1989. ISSN 0022-2364. doi: 10.1016/0022-2364(89)90011-5.

Kenneth P. Whittall, Alex L. Mackay, Douglas A. Graeb, Robert A. Nugent, David K. B. Li, and Donald W. Paty. In vivo measurement of T2 distributions and water contents in normal human brain. *Magnetic Resonance in Medicine*, 37(1):34–43, 1997. ISSN 1522-2594. doi: 10.1002/mrm.1910370107.

Alexander D. Wright, Michael Jarrett, Irene Vavasour, Elham Shahinfard, Shannon Kolind, Paul van Donkelaar, Jack Taunton, David Li, and Alexander Rauscher. Myelin Water Fraction Is Transiently Reduced after a Single Mild Traumatic Brain Injury – A Prospective Cohort Study in Collegiate Hockey Players. *PLOS ONE*, 11(2):e0150215, February 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0150215.

Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, and David Firmin. DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1310–1321, June 2018. ISSN 1558-254X. doi: 10.1109/TMI.2017.2785879.

A VARIANCE OF THE MMD

Let k be the reproducing kernel defining the MMD, and let $\mathbf{X}$ and $\mathbf{Y}$ be collections of m samples of $X, Y \in \mathbb{R}^n$. Following the notation in Sutherland (2019), let $\mathbf{K_{XY}} \in \mathbb{R}^{m \times m}$ be a matrix with elements $(\mathbf{K_{XY}})_{ij} = k(\mathbf{X}_i, \mathbf{Y}_j)$, with $\mathbf{K_{XX}}$ and $\mathbf{K_{YY}}$ defined similarly, and let $\tilde{\mathbf{K}}_{\mathbf{XX}}$ and $\tilde{\mathbf{K}}_{\mathbf{YY}}$ be equal to $\mathbf{K_{XX}}$ and $\mathbf{K_{YY}}$ with their diagonal elements set to zero. Then, equation (4) in Sutherland (2019) gives the unbiased estimator $\hat{V}_m(\mathbf{X}, \mathbf{Y})$ for $\operatorname{Var} \widehat{\mathrm{MMD}}^2_U(\mathbf{X}, \mathbf{Y})$ as

$$
\begin{aligned}
\hat{V}_m = & \frac{4}{(m)_4}\left[||\tilde{\mathbf{K}}_{\mathbf{XX}}\mathbf{1}||_2^2 + ||\tilde{\mathbf{K}}_{\mathbf{YY}}\mathbf{1}||_2^2\right] + \frac{4(m^2-m-1)}{(m)_2^3}\left[||\mathbf{K_{XY}}\mathbf{1}||_2^2 + ||\mathbf{K}_{\mathbf{XY}}^\top\mathbf{1}||_2^2\right] \\
& - \frac{8}{m(m)_3}\left[\mathbf{1}^\top\tilde{\mathbf{K}}_{\mathbf{XX}}\mathbf{K_{XY}}\mathbf{1} + \mathbf{1}^\top\tilde{\mathbf{K}}_{\mathbf{YY}}\mathbf{K}_{\mathbf{XY}}^\top\mathbf{1}\right] \\
& + \frac{8}{m^2(m)_3}\left[(\mathbf{1}^\top\tilde{\mathbf{K}}_{\mathbf{XX}}\mathbf{1} + \mathbf{1}^\top\tilde{\mathbf{K}}_{\mathbf{YY}}\mathbf{1})(\mathbf{1}^\top\mathbf{K_{XY}}\mathbf{1})\right] \\
& - \frac{2(2m-3)}{(m)_2(m)_4}\left[(\mathbf{1}^\top\tilde{\mathbf{K}}_{\mathbf{XX}}\mathbf{1})^2 + (\mathbf{1}^\top\tilde{\mathbf{K}}_{\mathbf{YY}}\mathbf{1})^2\right] - \frac{4(2m-3)}{(m)_2^3}(\mathbf{1}^\top\mathbf{K_{XY}}\mathbf{1})^2 \\
& - \frac{2}{(m)_4}\left[||\tilde{\mathbf{K}}_{\mathbf{XX}}||_F^2 + ||\tilde{\mathbf{K}}_{\mathbf{YY}}||_F^2\right] - \frac{4m(m-2)}{(m)_2^3}||\mathbf{K_{XY}}||_F^2
\end{aligned}
\tag{10}
$$

where $||\cdot||_2$ and $||\cdot||_F$ are the ℓ^2 and Frobenius norms, $\mathbf{1} \in \mathbb{R}^m$ is the vector of all ones, and $(m)_k \coloneqq m(m-1)\cdots(m-k+1)$ is the falling factorial.

Note that while verifying equation (4) in Sutherland (2019) using computer algebra, we found two typos (highlighted above). The first correction was to the denominator in the second term: $m^3(m-1)^2 \to m^3(m-1)^3 = (m)_2^3$. The second correction was a flipped sign on the last term: $+ \to -$. See the MATLAB script `mmd_variance.m` for verification.

B TRAINING

Model hyperparameters were optimized using a brute-force grid search across a broad range of parameter values. For training, we have access to a compute cluster with compute nodes containing 2.10 GHz Intel Xeon Gold 6130 processors with 16 CPU cores/32 threads each. Because of this infrastructure, we were able to train many models in parallel for each experiment. In all cases, models were trained for two days each using the Adam (Kingma & Ba, 2014) optimizer with default momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$; the optimizer step size η was varied.

B.1 GAN MODELS

Table 3: Hyperparameters tested for traditional GAN, MMD GAN, and hybrid-trained GAN adversarial models for both the toy and MRI experiments. Parameters for which multiple values were tested are listed within curly braces; the best model corresponds to the bolded value within the list.

	Toy GAN	Toy MMD	Toy Hybrid	MRI GAN	MRI MMD	MRI Hybrid
$N_{batches}$	10	10	10	10	10	10
m	$\{\mathbf{1024}, 2048\}$	2048	2048	$\{\mathbf{1024}, 2048\}$	2048	2048
$\eta/10^{-5}$	$\{\mathbf{10}, \sqrt{10}, 1\}$	$\{10\sqrt{10}, \mathbf{10}, \sqrt{10}\}$	$\{\mathbf{10}, \sqrt{10}\}$	$\{10, \mathbf{\sqrt{10}}, 1\}$	10	$\{10, \mathbf{\sqrt{10}}\}$
δ	0.1	0.1	0.1	0.025	$\{0.025, \mathbf{0.05}\}$	0.025
$[\log\epsilon^-, \log\epsilon^+]$	$[-8, -2]$	$[-8, -2]$	$[-8, -2]$	$[-6, -3]$	$[-6, -3]$	$[-6, -3]$
N_d	$\{64, \mathbf{128}\}$	128	128	$\{64, \mathbf{128}\}$	$\{\mathbf{128}, 256\}$	$\{\mathbf{128}, 256\}$
N_h	$\{2, \mathbf{4}\}$	4	4	$\{2, \mathbf{4}\}$	$\{2, \mathbf{4}\}$	4
D_{steps}	$\{1, 3, 5, 10, 15, \mathbf{20}\}$	–	$\{10, 15, \mathbf{20}\}$	$\{5, 10, 15, \mathbf{20}\}$	–	$\{10, \mathbf{15}, 20\}$
GAN_{rate}	–	–	$\{\mathbf{5}, 10, 25\}$	–	–	$\{5, 10, \mathbf{20}\}$
k loss	–	$\{\widehat{\mathrm{MMD}}^2_U, \hat{\boldsymbol{t}}\}$	$\{\widehat{\mathbf{MMD}}^{\mathbf{2}}_{\boldsymbol{U}}, \hat{t}\}$	–	$\{\widehat{\mathbf{MMD}}^{\mathbf{2}}_{\boldsymbol{U}}, \hat{t}\}$	$\{\widehat{\mathbf{MMD}}^{\mathbf{2}}_{\boldsymbol{U}}, \hat{t}\}$
k_{rate}	–	$\{10, 25, \mathbf{50}\}$	$\{10, 25, \mathbf{50}\}$	–	$\{3, 5, \mathbf{10}, 25\}$	$\{\mathbf{10}, 25\}$
N_{bw}	–	$\{\mathbf{4}, 8\}$	$\{\mathbf{4}, 8\}$	–	$\{2, 4, 6, \mathbf{8}\}$	$\{\mathbf{4}, 8\}$
$[\log\sigma^-, \log\sigma^+]$	–	$[-8, 4]$	$[-8, 4]$	–	$[-8, 4]$	$[-8, 4]$

Table 3 shows the hyperparameters which were tested for traditional GAN, MMD GAN, and hybrid-trained GAN models for both the toy and MRI experiments. During training, data sampled from the validation set was periodically used to perform maximum likelihood estimation (MLE) fits for each model. Following the training of all six models for all hyperparameter combinations, candidate top models were chosen for each of the six sweeps according to the highest log likelihood values (computed from a moving average over 25 epochs) resulting from the MLE fits on the validation data. Finally, each set of candidate top models were used for MLE fits to 1000 data samples from the held out testing data set for final model selection.

Figures 4 and 5 show example learned stochastic corrections for traditional, MMD, and hybrid GANs for the toy problem and MRI problem, respectively. In figure 4, we see an example of the traditional GAN learning a more uniform noise distribution for the toy problem – centered around the true value of $\epsilon_0 = 0.01$ – compared to the MMD GAN. This property appears to manifest in the hybrid GAN, as well. Figure 5 shows that, as with the toy problem, the three GANs for the MRI problem each learn similar stochastic corrections.

An extensive hyperparameter sweep was performed, but we would like to emphasize that this is not necessary to apply this method successfully. Rather, given that a sufficiently large computational infrastructure was available to us, we were interested in investigating the regions of hyperparameter space which resulted in GANs which either failed to converge, or were particularly successful. In the end, most parameters were not found to consistently affect the failure or success of the various GAN models, at least among the fairly conservative ranges of values which were tested. For example, all of the candidate top models were able to achieve the same optimized log likelihood values when accounting for the variance of the sample means. There were, however, some noteworthy parameters. The number of discriminator training steps, D_{steps}, for each generator training step for the traditional GAN models was 10 or higher in most successful models. We believe this is due to the baseline uncorrected model being already close to the data distribution, as illustrated in figure 2 of the main text, and therefore the discriminator D needs additional training steps to critique the subtle imperfections of the generator. The MMD kernel k, on the other hand, is a more powerful discriminator and had to be limited to being trained only every 10 or more epochs. Interestingly, training k via maximizing either of $\widehat{\mathrm{MMD}}_U^2$ or $\hat{t}$ performed similarly well. This is not surprising, however, as while optimizing $\hat{t}$ has theoretical advantages for increasing discriminator power, k is already in a sense "too powerful", and therefore we believe that making it even more powerful may not be helpful. Lastly, we experimented with increasing the size of the deterministic correction δ for the MRI MMD GAN. While the final best model did happen to use $\delta = 0.05$, this larger δ value did not consistently improve the models, and we preferred to keep the smaller $\delta = 0.025$ value for the other models. Note that in figure 5, we see that the learned distributions of deterministic corrections g_δ in the MRI experiment are nearly the same for traditional, MMD, and hybrid GANs, despite the larger δ used for the MMD GAN.

B.2 CVAE MODELS

Table 4: Hyperparameters tested for CVAE models for both the toy and MRI experiments. CVAE models are trained on $(\theta, \hat{Y})$ pairs where $\hat{Y} \sim G(X(\theta))$ are sampled from pre-trained MMD GAN generators G. Parameters for which multiple values were tested are listed within curly braces; the best model corresponds to the bolded value within the list. Both toy and MRI experiments swept over the same CVAE parameters.

	m	η	η_{min}	η_{rate}	η_{drop}	N_d	N_z	N_h
Toy CVAE	256	10^{-4}	10^{-5}	1000	$\{\mathbf{1}, \sqrt{10}\}$	$\{32, 64, \mathbf{128}\}$	$\{6, \mathbf{8}\}$	$\{2, \mathbf{4}, 6\}$
MRI CVAE	256	10^{-4}	10^{-5}	1000	$\{1, \sqrt{\mathbf{10}}\}$	$\{32, 64, \mathbf{128}\}$	$\{\mathbf{6}, 8\}$	$\{2, 4, \mathbf{6}\}$

The conditional variational autoencoder (CVAE) models use exactly the same architecture as described in Gabbard et al. (2019). Briefly, their CVAE architecture consists of three component networks: two encoders $\mathcal{E}_1 : \mathbb{R}^n \to \mathbb{R}^{2N_z}$ and $\mathcal{E}_2 : \mathbb{R}^{n+N_\theta} \to \mathbb{R}^{2N_z}$, and a decoder $\mathcal{D} : \mathbb{R}^{N_z} \to \mathbb{R}^{2N_\theta}$. $\mathcal{E}_1$ maps data samples $\hat{Y} \in \mathbb{R}^n$ to an N_z-dimensional latent space, where the $2N_z$ outputs of $\mathcal{E}_1$ parameterize N_z normal distributions; $\mathcal{E}_2$ maps $\theta \in \mathbb{R}^{N_\theta}$ and $\hat{Y} \in \mathbb{R}^n$ pairs to the same latent space; and $\mathcal{D}$ maps latent space samples $Z \in \mathbb{R}^{N_z}$ to $2N_\theta$ outputs which parameterize N_θ normal distribu-

tions. The normal distributions are parameterized by pairs $(\mu, \log \sigma)$ of means and log bandwidths. During training, the latent space encoding learned by $\mathcal{E}_1$, which is only given observations $\hat{Y}$, is informed by $\mathcal{E}_2$, which has access to both $\hat{Y}$ and underlying parameters θ. During inference, $\mathcal{E}_1$ samples are decoded by $\mathcal{D}$, producing approximate posterior distribution samples $\hat{\theta}$.

Table 4 shows the hyperparameters swept over during CVAE training for both the toy and MRI experiments, each model training on $(\theta, \hat{Y})$ pairs sampled from pre-trained MMD GAN generators $\hat{Y} \sim G(X(\theta))$. As in Gabbard et al. (2019), we use fully-connected networks for all three of $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{D}$, using ReLU activation functions on hidden layers and no activation function on output layers. We use the same notation for number of hidden nodes N_d and hidden layers N_h as in the GAN experiments, as well as for the initial step size η and batch size m. We additionally experimented with dropping the learning rate by a multiplicative factor η_{drop} every η_{rate} epochs, until a minimum value of η_{min}. CVAEs were robust to changes in hyperparameters, achieving nearly the same accuracy for any hyperparameter set. Candidate top CVAE models were first chosen through evaluating mean absolute error values of inferred $\hat{\theta}$ on the validation data set. The final best models were chosen through evaluation on a held out test data set.

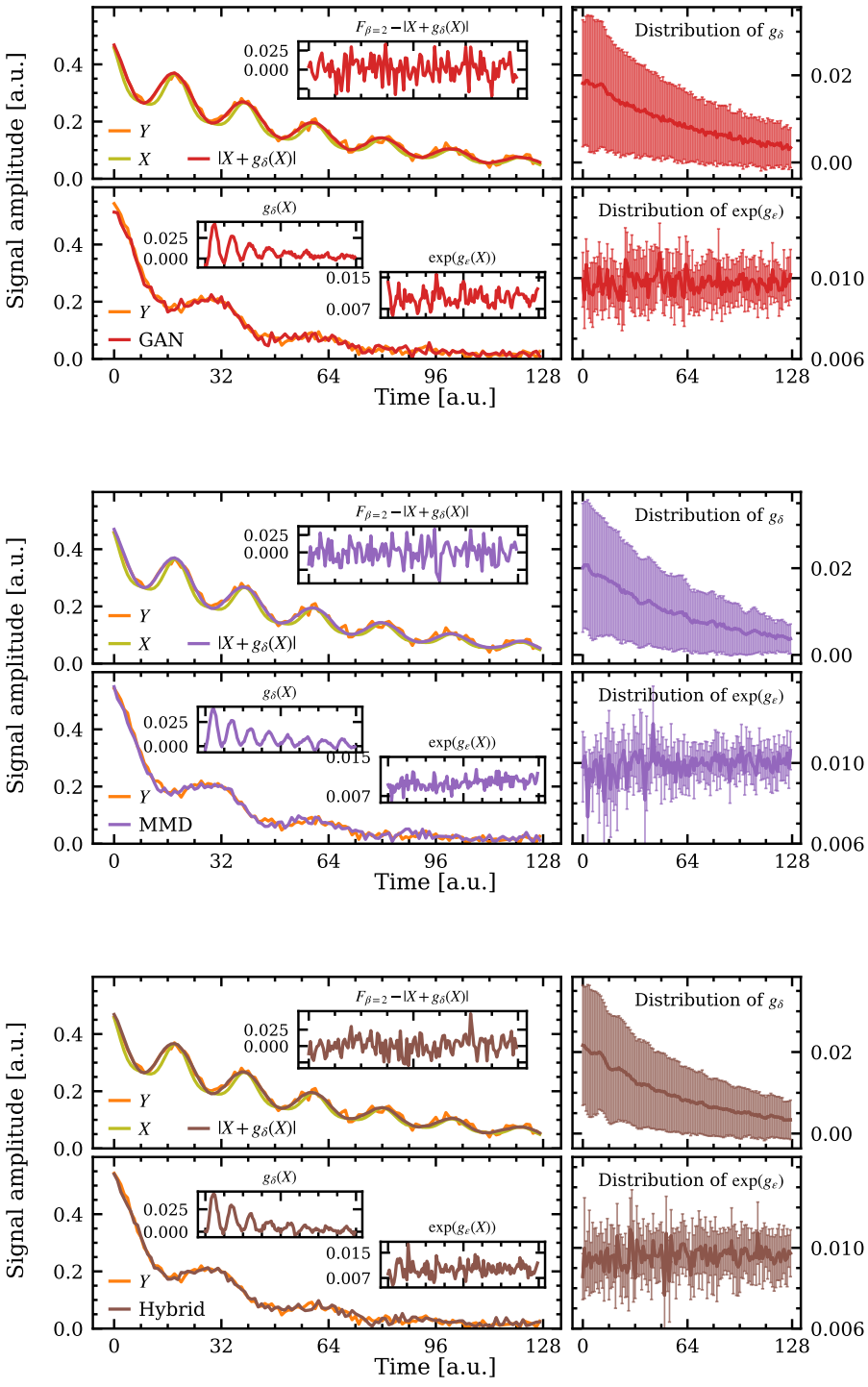

Figure 4: Example learned stochastic corrections for the toy problem using traditional, MMD, and hybrid GANs. The top row of plots for each model shows deterministic corrections, and the bottom row shows the stochastic corrections.

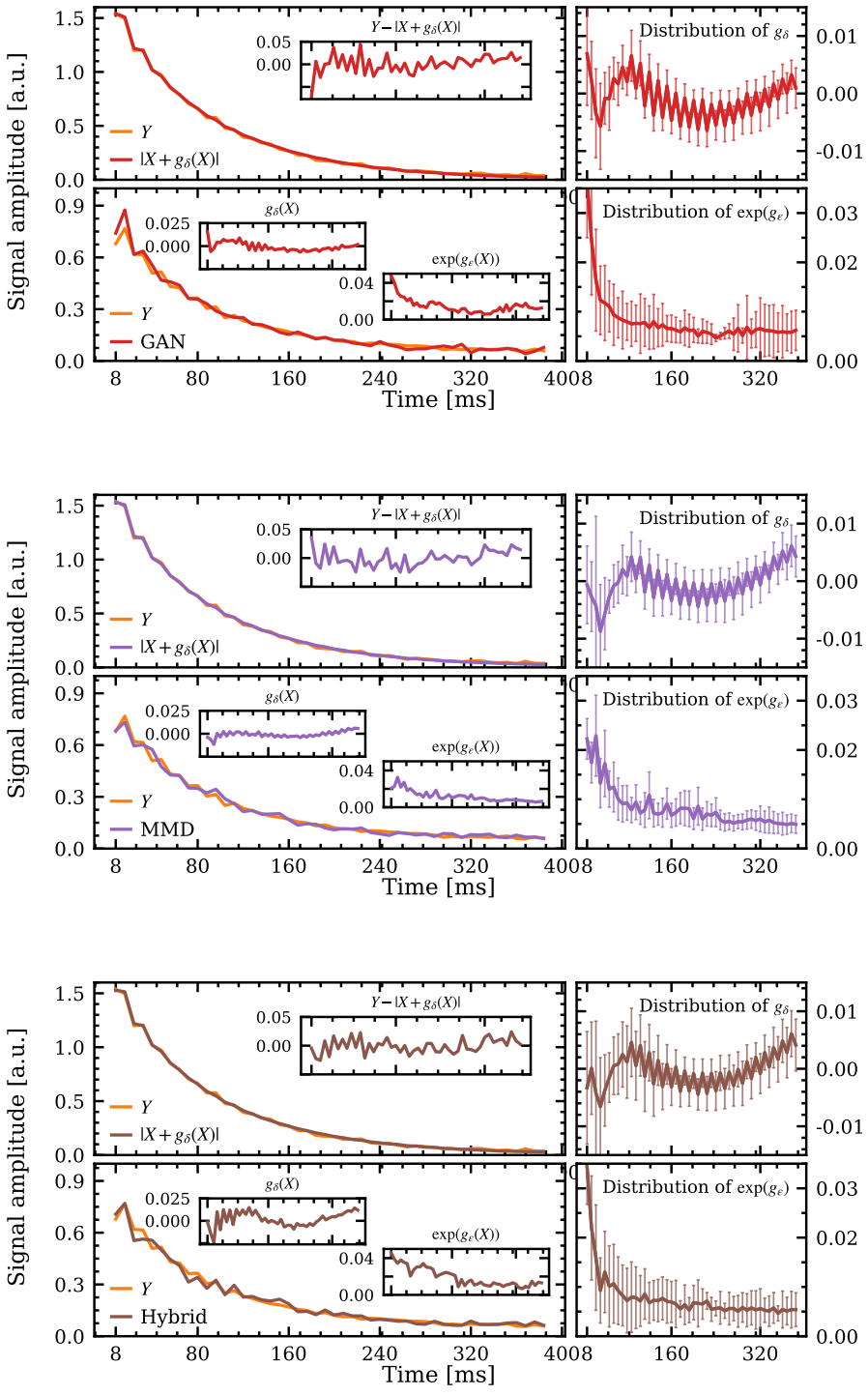

Figure 5: Example learned stochastic corrections for the MRI problem using traditional, MMD, and hybrid GANs. The top row of plots for each model shows deterministic corrections, and the bottom row shows the stochastic corrections.