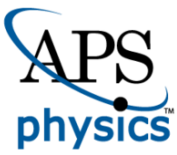

Reduced-Order Modeling with Artificial Neurons for Gravitational-Wave Inference

Alvin J. K. Chua, Chad R. Galley, and Michele Vallisneri

Phys. Rev. Lett. **122**, 211101 — Published 28 May 2019

DOI: 10.1103/PhysRevLett.122.211101

Reduced-order modeling with artificial neurons for gravitational-wave inference

Alvin J. K. Chua,* Chad R. Galley,† and Michele Vallisneri‡
Jet Propulsion Laboratory, California Institute of Technology,
4800 Oak Grove Drive, Pasadena, CA 91109, U.S.A.
(Dated: April 9, 2019)

Gravitational-wave data analysis is rapidly absorbing techniques from deep learning, with a focus on convolutional networks and related methods that treat noisy time series as images. We pursue an alternative approach, in which waveforms are first represented as weighted sums over reduced bases (reduced-order modeling); we then train artificial neural networks to map gravitational-wave source parameters into basis coefficients. Statistical inference proceeds directly in coefficient space, where it is theoretically straightforward and computationally efficient. The neural networks also provide analytic waveform derivatives, which are useful for gradient-based sampling schemes. We demonstrate fast and accurate coefficient interpolation for the case of a four-dimensional binary-inspiral waveform family, and discuss promising applications of our framework in parameter estimation.

PACS numbers: 02.50.Tt, 02.60.Gf, 04.30.-w, 04.80.Nn, 95.55.Ym

Introduction. Statistical inference is the eyepiece of gravitational-wave (GW) observatories: it maps the noise-dominated detector output into probabilistic assessments of candidate significance, and into posterior probability densities for the physical parameters of confirmed detections [1]. The mathematical setup of GW data analysis is simple, with most of its salient features apparent in the classical time-series likelihood

$$\mathcal{L}(\theta) := p(x|\theta) \propto \exp\left\{-\tfrac{1}{2}\left\langle x-h(\theta)|x-h(\theta)\right\rangle\right\}. \quad (1)$$

Here x is the detector output, h is the modeled detector response to an incoming GW with source parameters θ, and $\langle\cdot|\cdot\rangle$ is the (complex) noise-weighted inner product

$$\langle a(t)|b(t)\rangle := 4\int_0^\infty \mathrm{d}f\,\frac{\tilde{a}^*(f)\tilde{b}(f)}{S_n(f)}, \quad (2)$$

with tildes denoting Fourier transforms and S_n the one-sided power spectral density of detector noise n [2]. Eq. (1) essentially describes n as Gaussian and additive, with the sampling distribution $p(n) \propto \exp\{-\langle n|n\rangle^2/2\}$.

The challenges in using (1) for detection and parameter estimation are entirely practical. Detector physics poses issues such as noise characterization and response calibration; these are addressed by validating sample statistics with background studies, and including detector parameters during inference (e.g., [3, 4]). On the astrophysics side, the relativistic nature of many sources makes them computationally expensive to model, which hinders the bulk generation of accurate waveform templates for use in data-analysis algorithms. Furthermore, the nonlinear parameter dependence of such waveforms introduces complex features into the likelihood hypersurface (except at very high signal-to-noise ratios), making it difficult to map out with deterministic or stochastic methods.

In this Letter, we propose a general framework that combines order-reduction and machine-learning techniques to tackle these last two problems in unison, and to connect GW source modeling and data analysis in a natural and integrated manner. Our approach involves the construction of a fully representative reduced basis for the signal space of a GW model [5], and the fitting of a deep neural network to this parametrized manifold. The resultant function over parameter space is an analytic waveform in reduced representation; this enables the efficient generation of signal templates, and further allows the casting of the likelihood (1) in an equivalent reduced form. Also derived during the training of the network are the Jacobian and higher derivatives of the function, which encode the geometry of the signal manifold and can be used in derivative-based samplers for improved exploration of the likelihood hypersurface.

As a proof of principle, we present example results for a four-parameter post-Newtonian model of the GW signal from an inspiraling black-hole binary [6], which demonstrate the feasibility of our approach for higher-dimensional problems. A variety of network architectures are trialled to ensure the general premise is sound. The waveform error from fitted networks approaches (to within an order of magnitude) the error in reduced-order surrogate models [7–12], and several strategies for attaining even better accuracy are outlined. We showcase the speed and robustness of our networks on a number of derivative-based applications for parameter estimation, and discuss possible extensions of the framework.

Reduced-order modeling. Order-reduction strategies [13] are employed in GW source modeling and data analysis to represent waveform observables as linear combinations of reduced-basis vectors [5, 14]. Unlike standard transforms, the reduced-order modeling (ROM) approach is model-specific and involves building a finite, optimally compact basis whose span is essentially the full model space (to machine precision). This reduced basis is prepared offline (i.e., in advance of its use in data analysis) by means of a greedy algorithm [15] that distills a large set of training templates into a smaller orthonormal set of vectors. Any waveform in model space can be recon-

structed via projection onto the basis vectors, which are themselves linear combinations of training templates.

For a GW model $h(\theta)$ parametrized by $\theta \in \Theta \subset \mathbb{R}^s$, the ROM approach can be used to cast the signal space $\mathcal{S} := h[\Theta]$ in terms of a d-dimensional reduced basis $\{e_i\}$ with $\langle e_i|e_j\rangle = \delta_{ij}$,[1] such that $\mathcal{S}$ is isomorphic to an s-dimensional manifold embedded in $\mathbb{C}^d$. Through projection of the signal templates onto $\{e_i\}$, we may write

$$h(\theta) = \sum_i \langle h(\theta)|e_i\rangle e_i := \sum_i \alpha_i(\theta) e_i \equiv \alpha(\theta), \qquad (3)$$

with $\alpha \in \mathbb{C}^d$. The signal-to-noise ratio (SNR) for a template is given by $\rho := \langle h|h\rangle^{1/2} = (\alpha^\dagger\alpha)^{1/2} := |\alpha|$. It is convenient to work with normalized templates such that $|\alpha| = 1$; for some "true" signal $h(\theta_*)$ with arbitrary SNR ρ_*, we then have $h(\theta_*) \equiv \rho_*\alpha(\theta_*)$.

The computational efficiency of reduced-order surrogates [7–11] stems from the dimensionality d of the basis $\{e_i\}$ being small compared to the typical size r of the standard time-series representation h; the linearity of $\{e_i\}$ with respect to the inner product (2) can also be exploited to accelerate likelihood evaluations, through the method of reduced-order quadratures [16, 17]. However, all of these benefits rely on being able to obtain (for arbitrary source parameters θ) the basis projection coefficients $\alpha(\theta)$ without computing the full waveform $h(\theta)$ itself. Present applications make use of empirical interpolation [18], which requires h to be evaluated (or approximated) only at a set of d time/frequency nodes that are uniquely defined by the reduced basis. Direct interpolation of α across parameter space remains challenging, except in low-dimensional ($s \lesssim 3$) problems.

Artificial neural networks. The deep-learning paradigm encompasses a family of machine-learning techniques that automatically discover and process "features" in data, without the need for task-specific algorithms [19]. Most methods in deep learning are based on variants of artificial neural networks (ANNs) — biologically inspired computational objects comprising multiple layers of nonlinear processing nodes between the input to the network and its output. Within GW data analysis, ANNs trained under supervision (i.e., using data structured in the form of input–output pairs [20]) are potential "black-box" alternatives to more statistically principled methods based on Eq. (1), and they have recently been shown to achieve promising performance in detection and classification tasks [21–27].

ANNs are also well known to be universal approximators [28] for continuous functions: given an appropriate learning strategy and properly structured training data, a network of sufficient depth can interpolate between training examples to a high level of accuracy. In this Letter, we demonstrate that ANNs are in principle suited to the high-dimensional interpolation problem posed by the ROM projection coefficients. We design networks that take θ as their input, and output $\hat{\alpha}(\theta)$ (where the overhat is used here and henceforth to denote an interpolated estimate). These are fitted to a large set of training pairs $\{\theta_n, \alpha(\theta_n)\}$ chosen to adequately represent the domain of interest Θ; goodness of fit is then evaluated on a small test set of examples that are held out from the training stage. The result is a function $\hat{\alpha}(\theta) \approx \alpha(\theta)$ that is both fully analytic and computationally efficient, as it is composed of (many) closed-form array operations.

Convolutional-type network architectures [29] have become the dominant model in deep learning, due to high-profile successes in applications such as computer vision [30] and natural language processing [31]. They leverage spatial correlations in high-resolution data to lower the number of free parameters in the network, which in turn reduces overfitting. However, our investigations with deconvolutional neural networks [32] indicate that they are less suited to the ROM interpolation problem; here training data is abundant, and underfitting is more of an issue. This is because ROM coefficient data has already been pared down to its principal features, and in its present vector form cannot be organized spatially (although this might be possible with tensor decompositions/networks [33]). The low resolution ($d \sim 10^2$) of the data makes it more amenable to the "fully connected" layers in a multilayer perceptron [34], which are individually faster than convolutional layers at low width, and hence can be stacked to great depth for increased network capacity.

In our multilayer perceptron networks, the input layer θ and output layer $\hat{\alpha}$ are connected by a sequence of "hidden" layers, each parametrized by a matrix of weights w and a vector of biases b. The ℓ-th hidden layer takes an input $a_{\ell-1}$ from the previous layer, and outputs to the next layer the value $a_\ell = a(w_\ell a_{\ell-1} + b_\ell)$ of a closed-form, nonlinear "activation" function a. Values of the weights and biases that minimize a suitably defined loss of fidelity L in the output are learnt during the training stage, where the loss gradient $\partial L/\partial(w, b)$ is obtained through a backpropagation algorithm [35] and used iteratively in gradient descent optimization. ANNs can also be used to compute the analytic derivatives $\partial^n\hat{\alpha}/\partial\theta^n$ as functions of (w, b) (for $n-1$ up to the differentiability class of a); these converge to the target derivatives $\partial^n\alpha/\partial\theta^n$ with no added computational expense as the network is trained.

The reduced likelihood. We now consider the basic case where the data x in (1) is the sum of a single true signal $h(\theta_*)$ and the detector noise n. Projecting the data onto the reduced basis $\{e_i\}$ as in (3) gives $x \equiv \beta + \gamma$; here the reduced-space term $\beta = \rho_*\alpha(\theta_*) + \nu$ contains the true signal template $\alpha(\theta_*)$ and a noise component

[1] The reduced basis is orthonormal with respect to the specified noise power spectral density S_n in (1); for a different S_n such that $\langle e_i|e_j\rangle = N_{ij}$, pre-multiplying all coefficient vectors by the whitening matrix W (where $W^\dagger W = [N_{ij}]$ is obtained through, e.g., Cholesky decomposition) restores orthonormality.

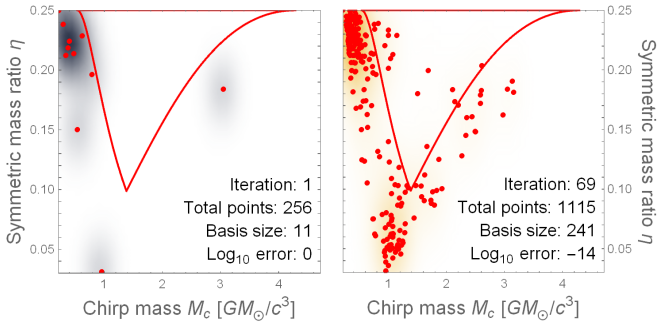

FIG. 1. Starting with a reduced basis built from 256 training points distributed uniformly over an extended domain, we draw up to 20 more training points from the kernel density estimate of the basis points at each iteration (red points, projected onto the (M_c, η)-plane), and construct a new basis. A final basis representation error of $\lesssim 10^{-14}$ over the domain of interest (red border) is attained with only $\sim 10^3$ training points, which is far more efficient than grid-based training.

$\nu \equiv \sum_i \langle n|e_i\rangle e_i$, while the orthogonal term $\gamma = x - \beta$ is the projection of n onto the orthogonal complement of S. In this representation, the likelihood (1) becomes

$$\begin{aligned}\mathcal{L}(\theta) &\propto \exp\left\{-\tfrac{1}{2}|\beta - \alpha(\theta)|^2 - \tfrac{1}{2}|\gamma|^2\right\},\\ &\propto \exp\left\{-\tfrac{1}{2}|\beta - \alpha(\theta)|^2\right\},\end{aligned} \tag{4}$$

since the orthogonal noise γ does not depend on θ.

Eq. (4) corresponds to the trivial probability model $\nu \sim \mathcal{N}(0, I_d)$ for detector noise in the reduced space $\mathbb{C}^d$, and retains the statistical properties of Eq. (1). Computation of the θ-dependent terms $\beta^\dagger\alpha(\theta)$ and $|\alpha(\theta)|^2$ in (4) is an $\mathcal{O}(d)$ operation, as compared to $\mathcal{O}(r \gg d)$ for $\langle x|h(\theta)\rangle$ and $\langle h(\theta)|h(\theta)\rangle$ in (1) (where r is the number of frequency-series samples).[2] With the efficient generation of $\hat{\alpha}(\theta)$ provided by a neural network, and the explicit reduction of complexity $r \to d$ in the reduced likelihood (4), much of the computational cost associated with online likelihood evaluations can thus be shifted into the offline construction of the reduced basis and ANN interpolant. If so desired, extrinsic parameters such as amplitude or time of arrival can still be handled analytically, as is customary in the GW literature [36].

Example results. In this work, we apply our approach to a 2.5PN TaylorF2 waveform family [6]. This analytic frequency-domain model describes the GW signal from an inspiraling black-hole binary with aligned spins, and is parametrized by the component masses $m_{1,2}$ and dimensionless spins $\chi_{1,2}$. We consider signals with $r = 10^4$ frequency samples over the range $[1, 10]$ mHz, and a parameter domain defined by $m_{1,2} \in [1.25, 10] \times 10^5 M_\odot$ and $\chi_{1,2} \in [-1, 1]$; these choices correspond to a subset of the high-redshift massive-black-hole binaries that will be observed by the space-based GW detector LISA [37].

Working in the more natural mass parametrization of chirp mass M_c and symmetric mass ratio η [38], we construct a reduced basis for the above model, but over an extended domain with $m_{1,2} \in [0.5, 15] \times 10^5 M_\odot$ and $\chi_{1,2} \in [-1.5, 1]$. This facilitates a new high-dimensional ROM training strategy [39] that iteratively builds up a kernel density estimate for the parameter-space distribution of templates selected by the greedy algorithm; in the present model, these tend to cluster at the low-mass and negative-spin boundaries (see Fig. 1). A subset of the extended domain is also used in the training of the ANNs, where compensating for the added structure helps to improve accuracy in the domain of interest. For the four-dimensional model, the size of the reduced basis is $d = 241$. A smaller basis is derived for the non-spinning two-dimensional submodel over the same range of masses.

Instead of interpolating $\alpha(\theta) \in \mathbb{C}^d \cong \mathbb{R}^{2d}$ with a single network, it is more practical to train two independent ANNs on its real and imaginary parts; this allows the layer width to be kept small, and it is straightforward to combine the two network outputs post hoc. We implement our ANNs using the standard `TensorFlow` software library [40]. Our network for the four-dimensional model contains 25 hidden layers a_ℓ, where the first five comprise $2^{\ell+2}$ nodes and the remaining layers have 256 nodes each. This choice of architecture is arrived at heuristically, with the maximal layer width determined by that of the output layer, and with 256-node layers being appended to a smaller initial network until a satisfactory level of accuracy is achieved (without overfitting).

While the ROM training set is far too sparse for interpolation purposes, it captures much of the underlying structure and might be used to inform the distribution of a larger ANN training set. However, a uniform grid of training examples over an extended domain can also yield good accuracy in the domain of interest. For the four-dimensional model, 6×10^5 training examples are needed to prevent overfitting, which we define as a difference of $> 0.1\%$ in median accuracy on the training and test sets. To enforce this, small validation sets of 5,000 examples are randomly generated throughout the training stage, and are used to inform the regularization method of early stopping [41]. Early stopping turns out to be unnecessary for the final network and full training set, since underfitting is present instead; this is indicated by a leveling-off in performance on both the training and validation sets as the network is trained.

The "leaky RELU" activation function [42] is applied on all hidden layers of the four-dimensional network, while linear activation (with a being the identity) is used on the output layer. Although the leaky RELU has good training efficiency [43], it is of class C^0 with vanishing second derivative, such that $\partial^2\hat{\alpha}/\partial\theta^2$ also vanishes glob-

[2] The reduced likelihood (4) shares the start-up cost of projecting the data with the reduced-order quadrature method [16, 17], but there the cost of the template–template term scales as $\mathcal{O}(d^2)$.

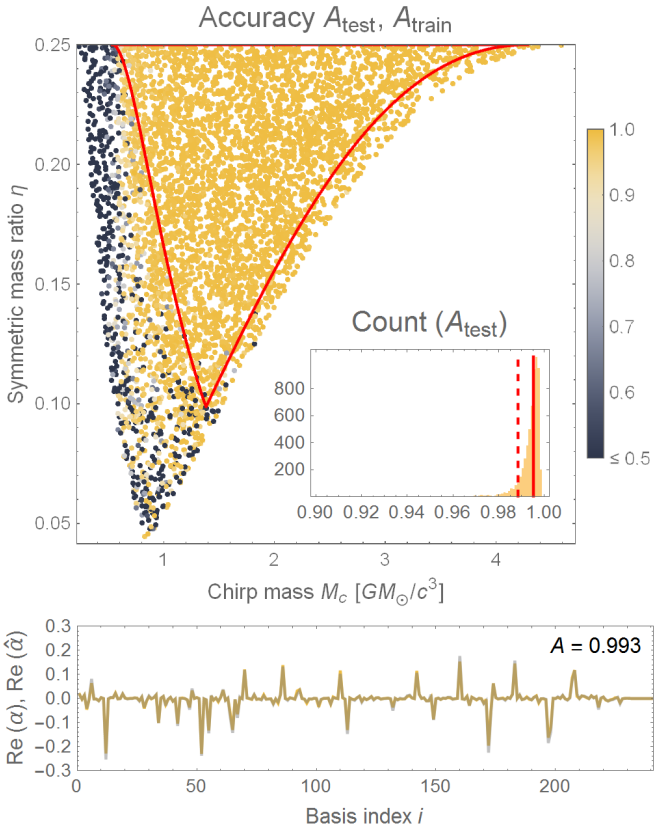

FIG. 2. Top: Plot of accuracy A as a function of (M_c, η) for test set (inside red border), and for 3,000 training examples with (M_c, η) outside the domain of interest. Inset: Histogram of test-set accuracy values with tenth percentile (dashed) and median (solid) indicated. Bottom: Visualization of typical coefficient vectors Re(α) (yellow) and Re($\hat{\alpha}$) (gray).

ally. This is not the case for the (slower-to-train) tanh activation function [44], which we employ in an ANN for the two-dimensional submodel. A mini-batch [19] size of 10^3 is empirically chosen for the training stage, and the adaptive, momentum-based ADAM optimization algorithm [45] (but with a manually decayed initial learning rate) is used to minimize the loss function[3]

$$L := \frac{\langle |\alpha - \hat{\alpha}|^2 \rangle}{\sqrt{\langle |\hat{\alpha}|^2 \rangle}}, \tag{5}$$

where $\langle \cdot \rangle$ denotes the average over a mini-batch.

To quantify the accuracy of $\hat{\alpha}$ on a test set of 5,000 examples, we use the normalized inner product (or overlap) between α and $\hat{\alpha}$, i.e., $A := \alpha^\dagger \hat{\alpha}/|\hat{\alpha}|$. The error $1 - A$ is related to (5) for a single template by $1 - A \leq L/2$, with equality in the case of an accurate norm ($|\hat{\alpha}| = 1$).

[3] This least-squares loss is weighted to converge from the direction of a larger norm, which helps to preserve the proportion between the independently trained real and imaginary parts.

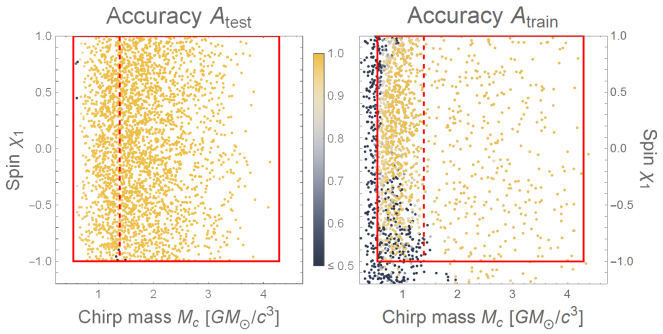

FIG. 3. Plot of accuracy A as a function of (M_c, χ_1) for test set (left), and for the 3,000 training examples in Fig. 2 (right).

Results for the four-dimensional model are presented in Figs 2 and 3. The ANN achieves a median accuracy of 99.5%; other high-dimensional ROM applications such as the numerical-relativity surrogates have typical median accuracies of 99.9% [9, 11]. Some of the disparity can be attributed to the four-fold larger mass ratio of eight in our model. Regardless, we estimate that the error in our ANN decays exponentially with network capacity: no more than 40 layers should be enough to hit 99.9%, which would only raise the network evaluation time from 1.5 to 2 ms (although the training-set size and the number of training epochs would have to be increased as well).

Parameter-estimation applications. The analytic waveform derivatives from our ANNs hold intriguing possibilities for GW data analysis, as they are immune to the speed and stability issues posed by taking derivatives numerically. For the above four-dimensional model, the Jacobian $J^a_i := \partial_i \hat{\alpha}^a$ is evaluated in ~ 0.1 s (without optimization); this yields fast and accurate estimates of the Fisher information matrix $F_{ij} := J_{ai} J^a_j$. The Fisher matrix describes the linearized (in $\hat{\alpha}$) likelihood around the maximum-likelihood estimate θ_{ML} [46], i.e.,

$$\mathcal{L}_1(\theta) \propto \exp\left\{-\tfrac{1}{2} F_{ij} \vartheta^i \vartheta^j\right\}, \tag{6}$$

where $\vartheta := \theta - \theta_{\text{ML}}$. Eq. (6) is inaccurate at low detection SNR, but may be improved with the Hessian $H^a_{ij} := \partial_i \partial_j \hat{\alpha}^a$. The noise-free second-order likelihood is [46, 47]

$$\mathcal{L}_2(\theta) \propto \mathcal{L}_1(\theta) \exp\left\{-\tfrac{1}{2} C_{ijk} \vartheta^i \vartheta^j \vartheta^k - \tfrac{1}{8} Q_{ijkl} \vartheta^i \vartheta^j \vartheta^k \vartheta^l\right\}, \tag{7}$$

where $C_{ijk} := J_{ai} H^a_{jk}$ and $Q_{ijkl} := H_{aij} H^a_{kl}$. In Fig. 4, we compare the probability contours of $\mathcal{L}_1$ and $\mathcal{L}_2$ to the stochastically sampled $\mathcal{L}$, for a low-SNR injected signal $\beta = 2\hat{\alpha}(\theta_*)$ in the two-dimensional (M_c, η)-submodel.

Another promising application is derivative-based sampling, which has hitherto been underutilized in GW parameter estimation due to the dearth of tractable waveform/likelihood derivatives. Most such techniques are Markov-chain Monte Carlo algorithms with gradient-informed chain dynamics (e.g., the Metropolis-adjusted Langevin algorithm [48, 49], Hamiltonian Monte Carlo

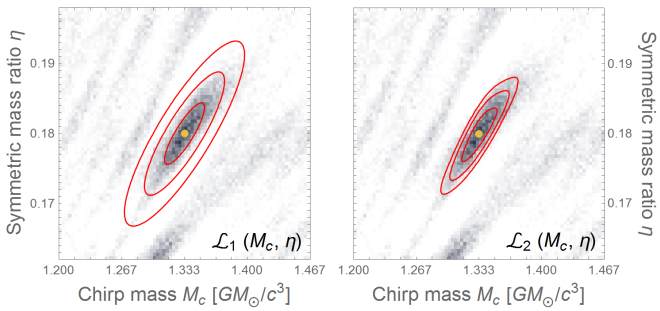

FIG. 4. One- to three-sigma contours (red) for $\mathcal{L}_1$ (left) and $\mathcal{L}_2$ (right) overlaid on density histogram of 10^5 samples drawn from $\mathcal{L}$, with $\theta_{\rm ML} = \theta_* = (4/3, 0.18)$ (yellow point).

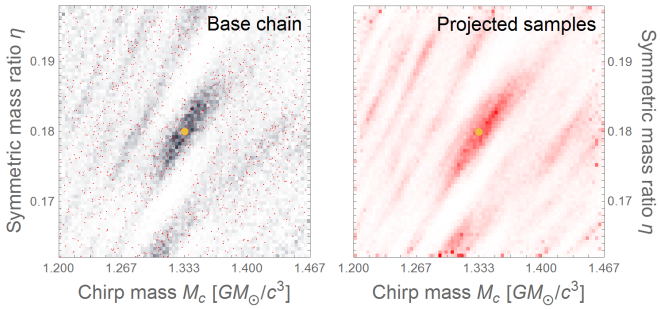

FIG. 5. Using the derivative-based sampling method of [52], a "base chain" (left; red points) is formed from the first 5,000 samples of the 10^5-sample chain in Fig. 4. Each base point seeds a "mini-distribution" of 20 points that are projected onto the tangent bundle of the signal manifold, and weighted accordingly in an approximate density histogram (right).

[50], and a family of stochastic-gradient variants [51]). They improve convergence near the maximum-likelihood point, but appear to be of less benefit in the more difficult global-search problem (due to many suppressed stationary points in the tail of a typical GW likelihood, which are present even at high SNR and can be found by mapping out the gradient field within our framework). A different type of derivative-based method for local sampling and density estimation exploits the constrained-Gaussian form of a target density such as Eq. (4) to produce approximate samples efficiently [52]. Applying this scheme to the above example gives the estimated histogram in Fig. 5, with order-of-magnitude computational savings.

Conclusion. We submit that ANNs are powerful tools for the high-dimensional interpolation of reduced-basis projection coefficients $\alpha(\theta)$, as necessary for the application of ROM to GW data analysis. Our approach is suitable for any waveform model whose signal space can be represented by a compact ($d \sim 10^2$) reduced basis; more extensive parameter domains may be dealt with piecewise. The ANNs provide fast, reliable derivatives that enable new techniques in GW parameter estimation. Another intriguing prospect is the possibility of inverting the ANN into a map from signal space to parameter space, in effect using ROM coefficients (obtained by projecting detector data onto the reduced basis) as natural machine-learning features. Such inverse ANNs could be trained on noisy data to provide quick maximum-likelihood estimates, supplemented by Fisher matrices from the forward map; their construction is left for future work.

Acknowledgements. We thank Natalia Korsakova and Michael Katz for helpful conversations, and we acknowledge feedback from fellow participants in the January 2018 LISA workshop at the Keck Institute for Space Studies. This work was supported by the JPL Research and Technology Development program, and was carried out at JPL, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

* alvin.j.chua@jpl.nasa.gov
† chad.r.galley@jpl.nasa.gov
‡ michele.vallisneri@jpl.nasa.gov

[1] B. F. Schutz. *Gravitational Wave Data Analysis.* Springer Netherlands, 2012.
[2] P. Jaranowski and A. Królak. *Analysis of Gravitational-wave Data.* Cambridge University Press, 2009.
[3] B. P. Abbott et al. GW150914: First results from the search for binary black hole coalescence with Advanced LIGO. *Phys. Rev. D*, 93(12):122003, June 2016.
[4] B. P. Abbott et al. Properties of the Binary Black Hole Merger GW150914. *Physical Review Letters*, 116(24):241102, June 2016.
[5] S. E. Field, C. R. Galley, F. Herrmann, J. S. Hesthaven, E. Ochsner, and M. Tiglio. Reduced Basis Catalogs for Gravitational Wave Templates. *Phys. Rev. Lett.*, 106:221102, Jun 2011.
[6] K. G. Arun, A. Buonanno, G. Faye, and E. Ochsner. Higher-order spin effects in the amplitude and phase of gravitational waveforms emitted by inspiraling compact binaries: Ready-to-use gravitational waveforms. *Phys. Rev. D*, 79:104023, May 2009.
[7] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio. Fast Prediction and Evaluation of Gravitational Waveforms Using Surrogate Models. *Phys. Rev. X*, 4:031006, Jul 2014.
[8] J. Blackman, S. E. Field, C. R. Galley, B. Szilágyi, M. A. Scheel, M. Tiglio, and D. A. Hemberger. Fast and Accurate Prediction of Numerical Relativity Waveforms from Binary Black Hole Coalescences Using Surrogate Models. *Phys. Rev. Lett.*, 115:121102, Sep 2015.
[9] J. Blackman, S. E. Field, M. A. Scheel, C. R. Galley, D. A. Hemberger, P. Schmidt, and R. Smith. A surrogate model of gravitational waveforms from numerical relativity simulations of precessing binary black hole mergers. *Phys. Rev. D*, 95:104023, May 2017.
[10] B. D. Lackey, S. Bernuzzi, C. R. Galley, J. Meidam, and C. Van Den Broeck. Effective-one-body waveforms for binary neutron stars using surrogate models. *Phys. Rev. D*, 95:104036, May 2017.

[11] J. Blackman, S. E. Field, M. A. Scheel, C. R. Galley, C. D. Ott, M. Boyle, L. E. Kidder, H. P. Pfeiffer, and B. Szilágyi. Numerical relativity waveform surrogate model for generically precessing binary black hole mergers. *Phys. Rev. D*, 96:024058, Jul 2017.

[12] Vijay Varma, Scott E. Field, Mark A. Scheel, Jonathan Blackman, Lawrence E. Kidder, and Harald P. Pfeiffer. Surrogate model of hybridized numerical relativity binary black hole waveforms. *arXiv e-prints*, page arXiv:1812.07865, December 2018.

[13] W. Schilders, H. Van der Vorst, and J. Rommes. *Model Order Reduction: Theory, Research Aspects and Applications*, volume 13. Springer, 2008.

[14] J. Blackman, B. Szilagyi, C. R. Galley, and M. Tiglio. Sparse Representations of Gravitational Waves from Precessing Compact Binaries. *Phys. Rev. Lett.*, 113:021101, Jul 2014.

[15] V. N. Temlyakov. Greedy approximation. *Acta Numerica*, 17:235409, 2008.

[16] P. Cañizares, S. E. Field, J. R. Gair, and M. Tiglio. Gravitational wave parameter estimation with compressed likelihood evaluations. *Phys. Rev. D*, 87:124005, 2013.

[17] P. Cañizares, S. E. Field, J. R. Gair, V. Raymond, R. Smith, and M. Tiglio. Accelerated Gravitational Wave Parameter Estimation with Reduced Order Modeling. *Phys. Rev. Lett.*, 114:071104, Feb 2015.

[18] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus Mathematique*, 339(9):667–672, 2004.

[19] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[20] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.

[21] T. Gebhard, N. Kilbertus, G. Parascandolo, I. Harry, and B. Schölkopf. ConvWave: Searching for Gravitational Waves with Fully Convolutional Neural Nets. In *Workshop on Deep Learning for Physical Sciences (DLPS) at the 31st Conference on Neural Information Processing Systems (NIPS)*, December 2017.

[22] D. George and E. A. Huerta. Deep neural networks to enable real-time multimessenger astrophysics. *Phys. Rev. D*, 97:044039, Feb 2018.

[23] D. George and E.A. Huerta. Deep Learning for real-time gravitational wave detection and parameter estimation: Results with Advanced LIGO data. *Physics Letters B*, 778:64 – 70, 2018.

[24] M. Razzano and E. Cuoco. Image-based deep learning for classification of noise transients in gravitational wave detectors. *Class. Quantum Grav.*, 35(9):095016, 2018.

[25] H. Gabbard, M. Williams, F. Hayes, and C. Messenger. Matching Matched Filtering with Deep Networks for Gravitational-Wave Astronomy. *Phys. Rev. Lett.*, 120:141103, Apr 2018.

[26] D. George, H. Shen, and E. A. Huerta. Classification and unsupervised clustering of LIGO data with Deep Transfer Learning. *Phys. Rev. D*, 97:101501, May 2018.

[27] A. Rebei, E. A. Huerta, S. Wang, S. Habib, R. Haas, D. Johnson, and D. George. Fusing numerical relativity and deep learning to detect higher-order multipole waveforms from eccentric binary black hole mergers. *ArXiv e-prints*, July 2018.

[28] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989.

[29] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 255–258. MIT Press, Cambridge, MA, USA, 1998.

[30] N. Sebe, I. Cohen, A. Garg, and T. S. Huang. *Machine Learning in Computer Vision*. Springer Netherlands, 2005.

[31] Y. Goldberg. A Primer on Neural Network Models for Natural Language Processing. *J. Artif. Int. Res.*, 57(1):345–420, September 2016.

[32] R. Fergus, M. D. Zeiler, G. W. Taylor, and D. Krishnan. Deconvolutional networks. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)*, volume 00, pages 2528–2535, 06 2010.

[33] A. Cichocki. Era of Big Data Processing: A New Approach via Tensor Networks and Tensor Decompositions. *ArXiv e-prints*, March 2014.

[34] S. S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1999.

[35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[36] A. Buonanno, Y. Chen, and M. Vallisneri. Detection template families for gravitational waves from the final stages of binary black-hole inspirals: Nonspinning case. *Phys. Rev. D*, 67(2):024016, January 2003.

[37] K. Danzmann et al. Laser Interferometer Space Antenna. *ArXiv e-prints*, February 2017.

[38] C. J. Cutler and É. E. Flanagan. Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral waveform? *Phys. Rev. D*, 49:2658–2697, Mar 1994.

[39] C. R. Galley. In prep.

[40] M. Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015.

[41] S. Amari, N. Murata, K. . Muller, M. Finke, and H. H. Yang. Asymptotic statistical theory of overtraining and cross-validation. *IEEE Transactions on Neural Networks*, 8(5):985–996, Sep. 1997.

[42] A. Y. Ng A. L. Maas, A. Y. Hannun. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[43] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, and M. Dudk, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Apr 2011.

[44] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient BackProp. In G. Montavon, G. B. Orr, and K.-R. Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[45] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *ArXiv e-prints*, December 2014.

[46] M. Vallisneri. Use and abuse of the Fisher information matrix in the assessment of gravitational-

wave parameter-estimation prospects. *Phys. Rev. D*, 77:042001, Feb 2008.

[47] E. Sellentin, M. Quartin, and L. Amendola. Breaking the spell of Gaussianity: forecasting with higher order Fisher matrices. *Monthly Notices of the Royal Astronomical Society*, 441(2):1831–1840, 2014.

[48] G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[49] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[50] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216 – 222, 1987.

[51] Y.-A. Ma, T. Chen, and E. B. Fox. A Complete Recipe for Stochastic Gradient MCMC. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2917–2925, Cambridge, MA, USA, 2015. MIT Press.

[52] Alvin J. K. Chua. Sampling from manifold-restricted distributions using tangent bundle projections. *arXiv e-prints*, page arXiv:1811.05494, November 2018.