LASER INTERFEROMETER GRAVITATIONAL WAVE OBSERVATORY
- LIGO -
CALIFORNIA INSTITUTE OF TECHNOLOGY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Technical Note **LIGO-T2000181-v2** 2020/05/22

Nonlinear Regression of Technical Noise in LIGO with Machine Learning

Hang Yu, Zachary Mark, Yuntao Bai, Rana X Adhikari
Division of Physics, Math, and Astronomy,
California Institute of Technology

California Institute of Technology
LIGO Project, MS 100-36
Pasadena, CA 91125
Phone (626) 395-2129
Fax (626) 304-9834
E-mail: info@ligo.caltech.edu

Massachusetts Institute of Technology
LIGO Project, Room NW22-295
Cambridge, MA 02139
Phone (617) 253-4824
Fax (617) 253-7014
E-mail: info@ligo.mit.edu

WWW: http://www.ligo.caltech.edu/

Contents

1 Introduction **3**

2 Examples of Nonlinear Coupling Mechanisms **3**

3 Regression with Neural Networks **6**

A Astrophysical Motivations **7**

B Examples of Current Efforts **7**

B.1 Bilinear Coupling 11

B.2 Realistic angle-to-length noise 11

B.3 Obstacles 11

1 Introduction

Broadly speaking, the sensitivity of aLIGO is determined by two kinds of noise sources, an unsubtractable one and a subtractable one. The former originates from quantum or thermal fluctuations and shows up only in the main GW readout channel. It cannot be distinguished from the GW signal and thus sets a fundamental limit of the instrument's sensitivity. On the other hand, the latter one is due to cross-couplings from the auxiliary control loops and/or some environmental perturbations like the seismic motion. The perturbations causing excess noises in DARM are also continuously recorded in (tens of thousands of) auxiliary channels, and therefore can in principle be used to reconstruct and then clean up the contamination showing up in the main GW readout.

In Figure 1 we show the noise budget of aLIGO in its first observing run [1]. As can be seen from the plot, aLIGO's sensitivity does not reach its fundamental limit set by the quantum and thermal noises until 100 Hz. In fact, the total noise (the red "Measured Noise" trace in Figure 1) in the $10-20$ Hz band the noise is nearly two orders of magnitude above its fundamental limit determined by the quantum (the grey "Quantum Noise" trace) and thermal noise (the blue "Thermal Noise" trace).

Our goal, as illustrated in the flow diagram in Figure 2, is to develop machine-learning-based nonlinear regression techniques to remove the auxiliary channels' contamination to the GW channel, and hence improve the aLIGO sensitivity in the sub-100 Hz band.

2 Examples of Nonlinear Coupling Mechanisms

How a noise source propagates to the main GW readout is often a complicated (and sometimes unknown) process. We show in Figure 3 two examples of typical nonlinear noise coupling that happens in aLIGO.

In the left panel, the angular motion of the mirror and the beam spot motion can couple to create a length signal that mimics the GW. If the beam spot is displaced Δy from the rotational pivot and the mirror is rotated by $\Delta\theta$, it creates a fluctuation in length, ΔL,

$$\Delta L(t) = \Delta y(t)\Delta\theta(t) \tag{1}$$

where the quantities are defined in Figure 3a. In this case, the contamination is the product of two auxillary noise sources (LIGO doesn't directly measure Δy but it can be inferred from other channels).

Another important noise is due to backscattering, which is illustrated in the right panel. Because of defects of a mirror's surface, it can scatter off some light from the main beam (in Figure 3b this occurs at the end test mass EX, but it can also occur at other optics). The stray lights may reflect upon some scattering objects (e.g., chamber walls) and recombined to the main beam. This process creates light fields ΔE whose phase is shifted with respect to the main field E_0 by an amount of

$$\frac{\Delta E}{E_0}(t) \propto \exp\left[4\pi \mathrm{i}\frac{\Delta x(t)}{\lambda}\right], \tag{2}$$

where λ is the laser's wavelength and $\Delta x(t)$ is the relative displacement between the mirror and scattering objects. When $\Delta x(t) \gtrsim \lambda$, the scattered field $\Delta E(t)$ becomes nonlinear and can up-scatter the large, low-frequency seismic motion into the band of GW readout.

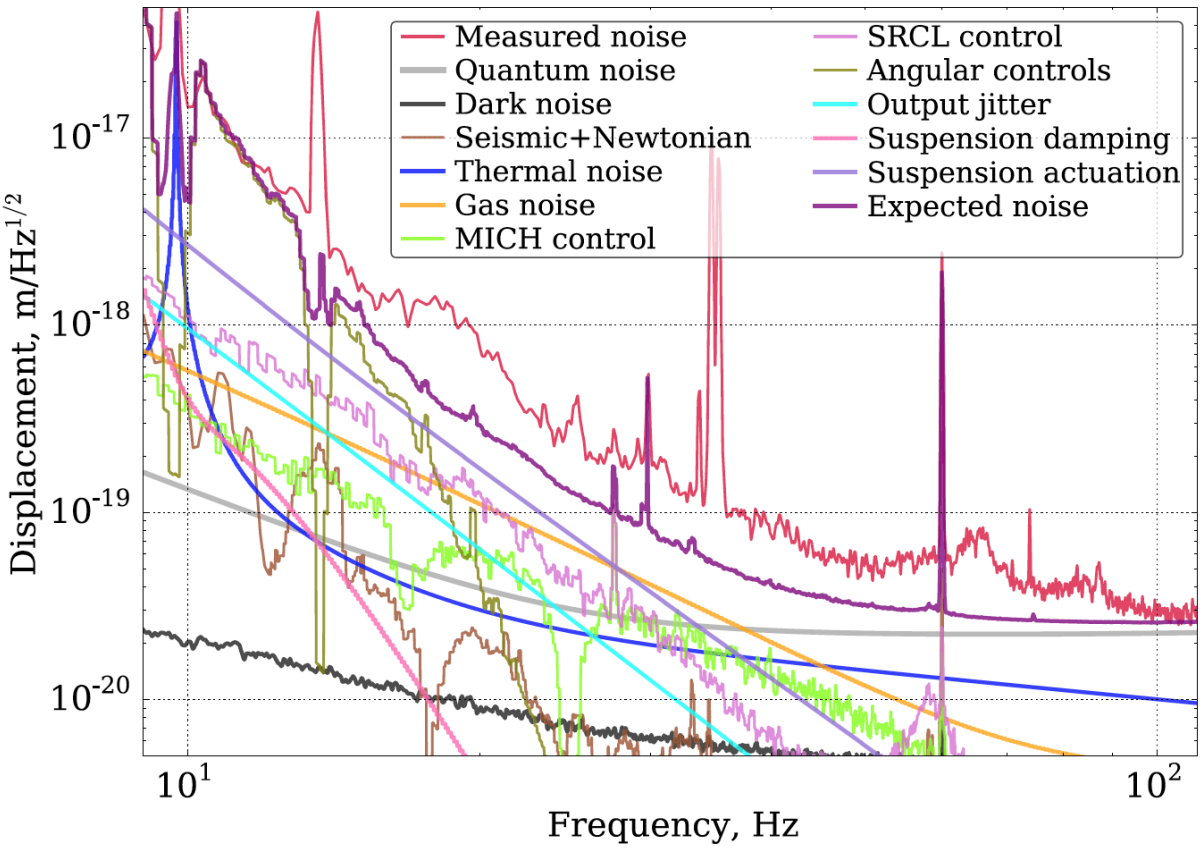


FIGURE 1: Noise budget of aLIGO [1]. In the 10 – 100 Hz band, the sensitivity is limited by predictable noises (i.e., noises that are not due to quantum or thermal fluctuations) and thus indicates a large room of potential improvement. The 'Expected Noise' trace is the quadrature sum of all budgeted noise sources and 'Measure noise' is the actual output of the LIGO detector.

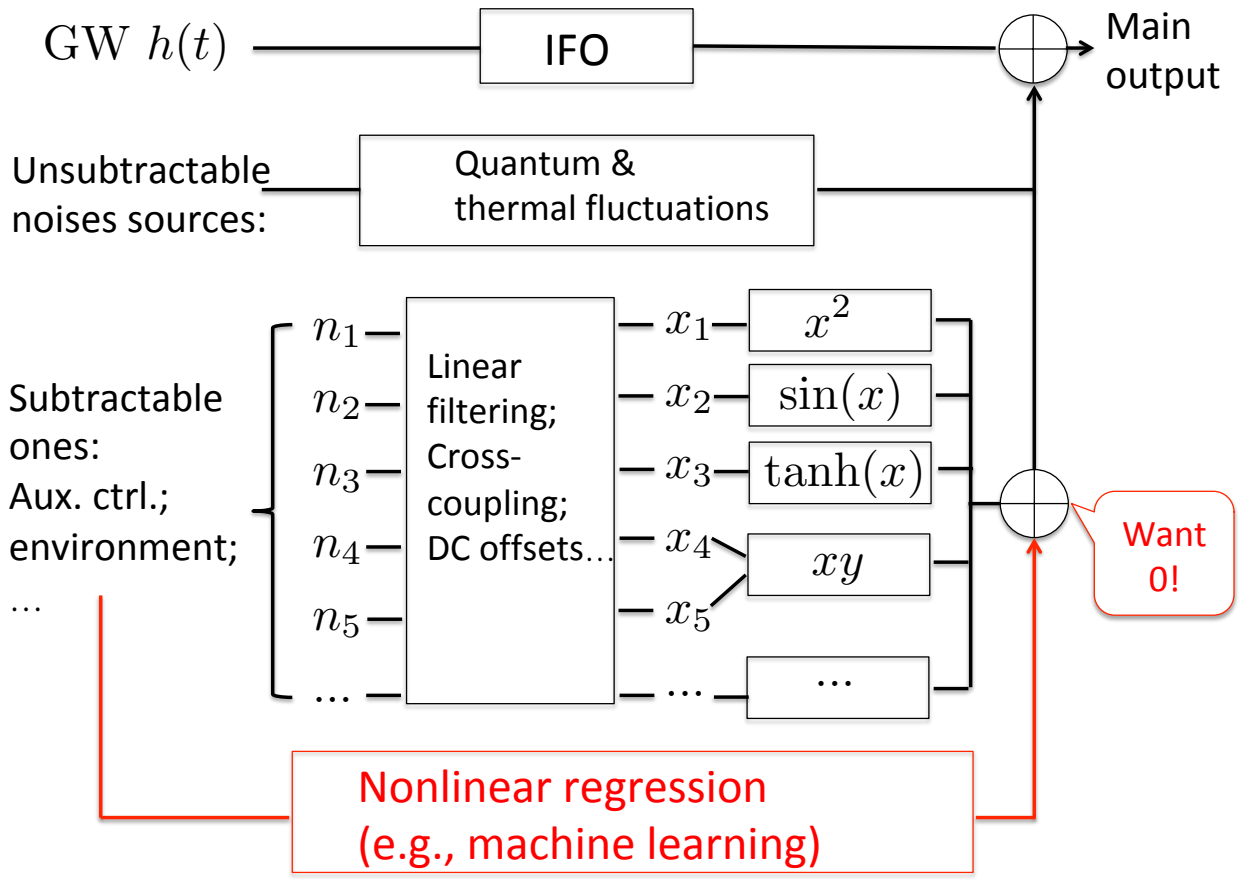
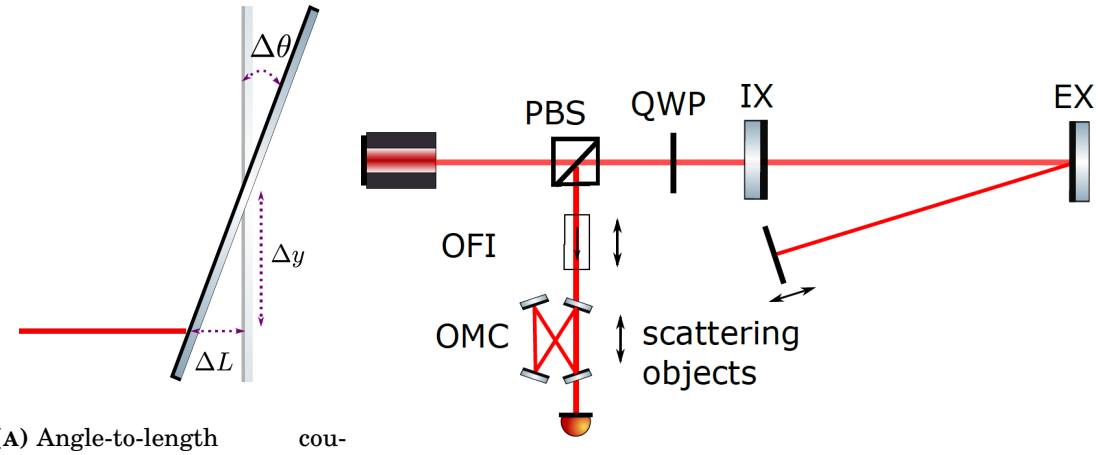


FIGURE 2: Flow diagram of signal and noise propagation. While the original perturbations of the subtractable noises are recorded, identifying their couplings to the main GW readout is challenging. In part, this is due to the large number of channels involved (thousands to tens of thousands). More importantly, the couplings often have both linear and nonlinear components, and hence cannot be removed with classical linear regression methods. Instead, our goal is to develop machine-learning-based nonlinear regression techniques to use time series from the auxiliary channels to construct the disturbance and then remove it from the main GW readout.



(A) Angle-to-length coupling [2].

(B) Scattering noise coupling [3].

FIGURE 3: Examples of nonlinear noise coupling in LIGO.

3 Regression with Neural Networks

Before tackling real data, we simplify the problem by working with mock data sets that mimic the real interferometer. Each of our mock data sets consists of:

- **A target channel**, which is a time-series modeling simulated noise in the GW readout, also called darm in the context of interferometers.
- **Witness channels**, which are the time series necessary to predict the target channel.

We increase the complexity of the mock data sets as we have success. Some of the complications built into are mock data sets include:

- **Imperfect witnesses.** Certain quantities like the beam's location on a mirror or the motion of the exact scattering point (as we don't know its exact location), are often not directly measured. Instead, they themselves need to be first inferred from other channels (e.g., the spot location on a mirror may be contained in a combination of seismometers and angular sensors; the motion of the unknown scattering point can be interpolated from a sets of accelerometers nearby).
- **Frequency dependent filtering.** The auxillary channels (which are measured in digital counts) often require linear, frequency-dependent filtering to become physically relevant quantities like the longitudinal or angular motion of the mirror.

We try to find a neural network that can efficiently predict the target channel from the witness channels for all of our mock data sets. So far we have found that

- 1D convolutional neural networks perform well, especially for learning phase shifts and frequency-dependent filtering.
- Dense layers with nonlinear activation functions are useful for learning nonlinear couplings between channels
- Dropout layers are necessary to prevent overfitting.

Some specific examples are shown in Appendix A.

A Astrophysical Motivations

Astrophysically, removing noise in the sub$-100\,$Hz band will have significant outcomes.

1. Early warning of binary meutron star mergers
2. Higher Mass binary black holes
3. cosmology with high redshift sources
4. BBH astrophysical foreground

For example, the gravitational-wave signal of a black hole binary has a characteristic frequency scale given by

$$f_{\rm merger} \simeq 40 \left(\frac{3}{1+z}\right)\left(\frac{100\,M_\odot}{M_{\rm tot}}\right) {\rm Hz}, \tag{3}$$

where z is the cosmological redshift and $M_{\rm tot}$ is the total mass of the system. Therefore, in order to detect massive binaries at high cosmological redshifts, it is critical to improve the low-frequency sensitivity that is limited by predictable noises.

More quantitatively, we note that

$${\rm SNR}^2 = 4\int \frac{f h^*(f) h(f)}{S_{\rm n}(f)} {\rm d}\log f, \tag{4}$$

where $h(f)$ is the frequency-domain GW waveform and $S_{\rm n}(f)$ is the power spectral density of the noise. We can hence define

$$\rho^2 \equiv 4\frac{f|h(f)|^2}{S_{\rm n}(f)}, \tag{5}$$

as a density that measures the contribution to the total signal-to-noise ratio (SNR) per $\log f$.

We show this quantity in Appendix A for three different detector sensitivities: O2 (blue), the design sensitivity of aLIGO (orange), and A+ (green). Here we consider a system with total mass of $(1+z)M_{\rm tot} = 300\,M_\odot$ in the detector frame. In order to emphasize the contribution from each $\log f$ interval, we have normalized each curve by its peak value. As shown in the figure, the excess (yet subtractable) low-frequency noise reduces the SNR by a large amount. If our machine learning technique can eventually clean the low-frequency region, it could enhance the total SNR by 40%, which means amplifying the search volume of such systems by a factor of $\sim 1.4^3 \simeq 2.7$.

We also show the detection horizon as a function of the detector-frame total mass in Fig. 5. In order to detect systems with $(1+z)M_{\rm tot} \gtrsim 300\,M_\odot$, (which can be intrinsically massive systems involving intermediate-mass black holes or can be systems at high cosmological redshift), it is again critical to improve the low-frequency sensitivity.

B Examples of Current Efforts

We implement our networks with Keras [4] in Python.

Here we show results obtained with a network consisting of a convolutional layer with a linear activation function, followed by alternating pairs of dropout and dense layers with

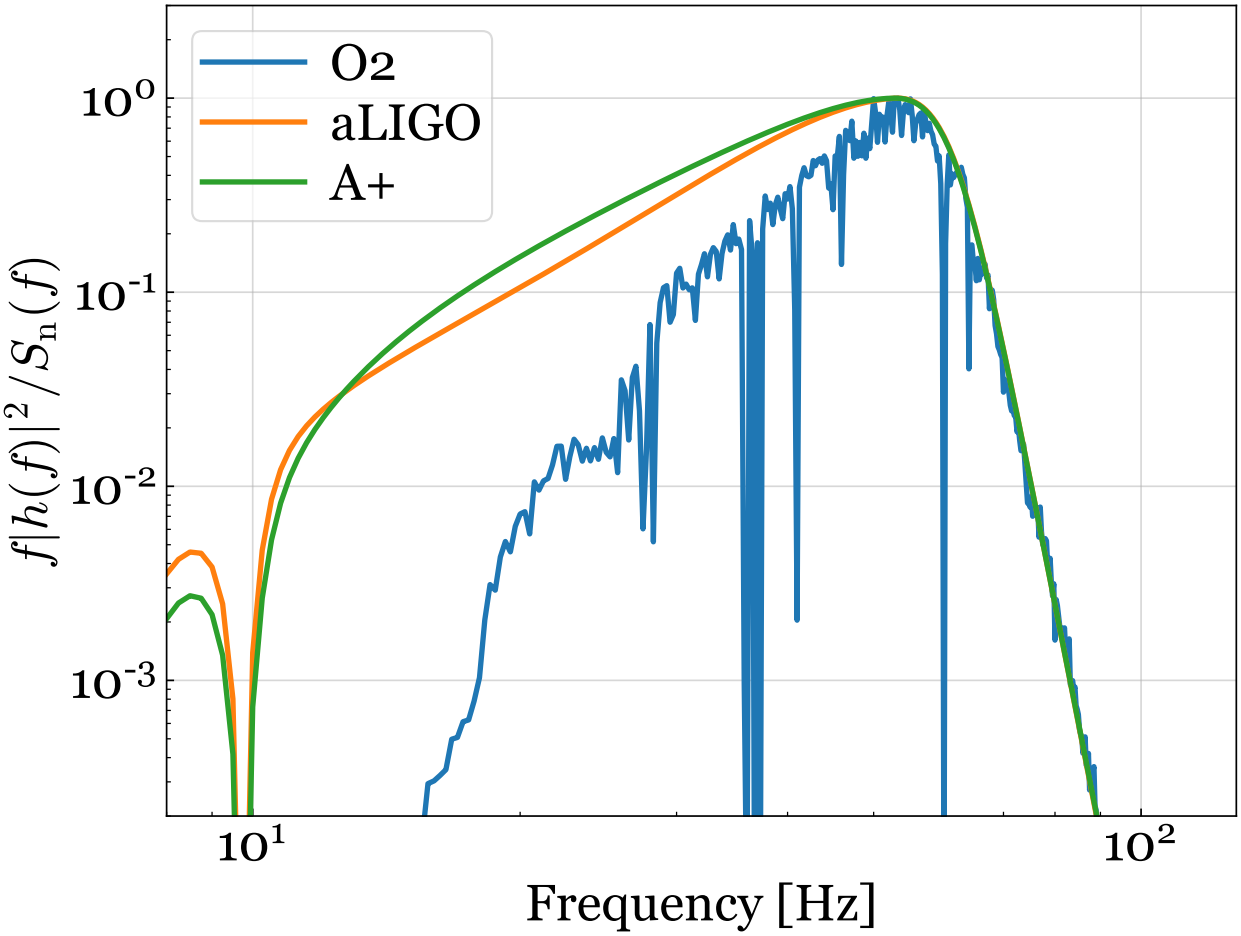


FIGURE 4: Density of square SNR as a function of frequency for a system with detector-frame total mass of $(1+z)M_{\rm tot} = 300\,M_\odot$. Each curve has been normalized by its peak value. Compared to the design sensitivity, we are losing a large amount of SNR due to the excess low-frequency noise.

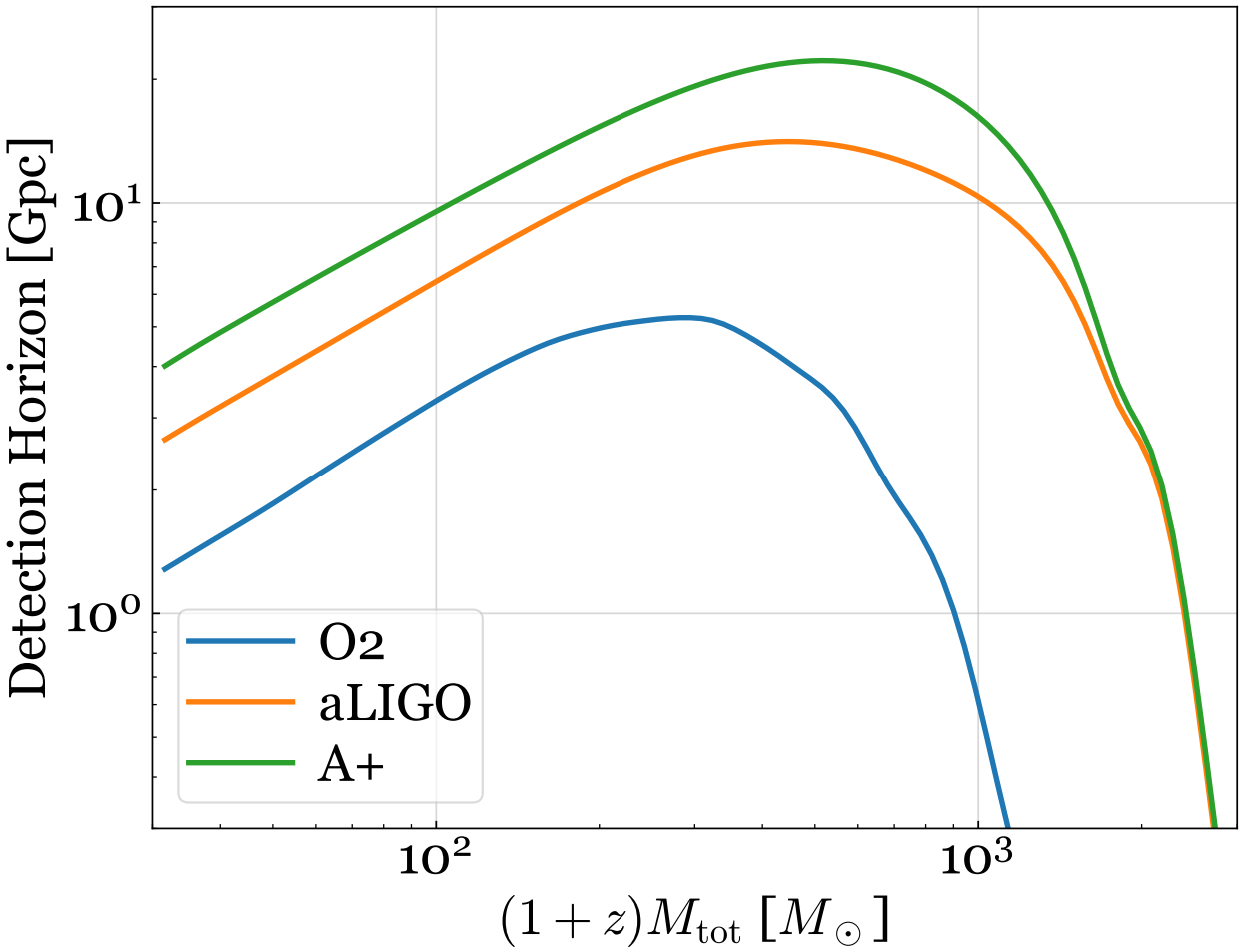


FIGURE 5: Detection horizon.

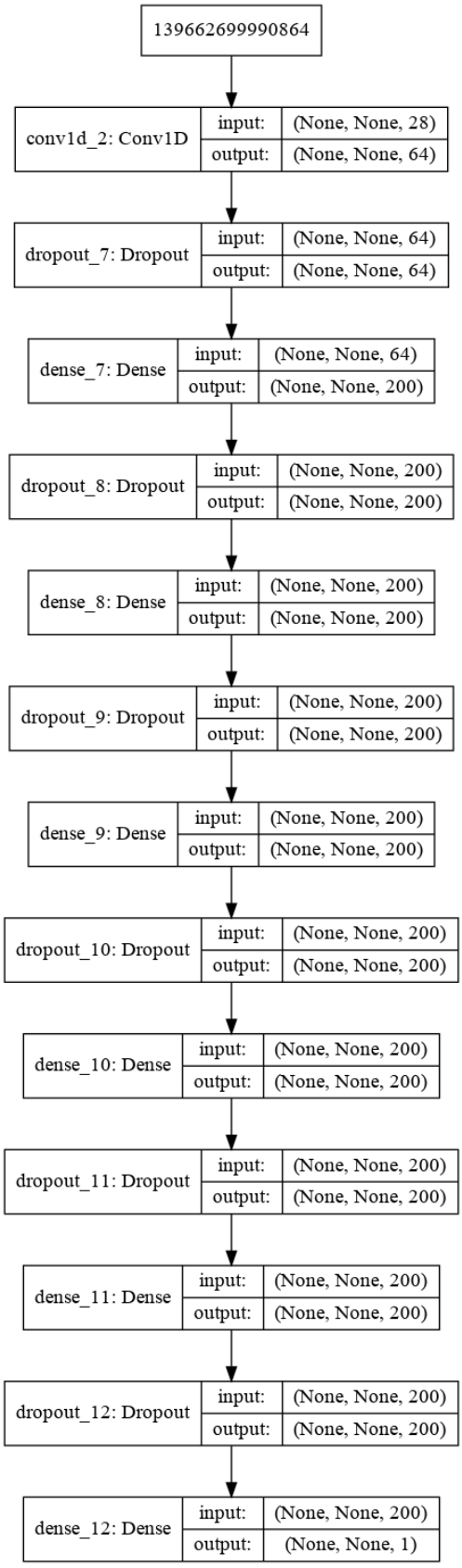


FIGURE 6: The CNN-Dense neural network used in these examples.

eLU activations. The examples here use 4 dropout-dense pairs. The final dense layer only has one neuron so that the output is a time series just like the target channel. This network is visualized in Figure 6.

We are also looking into replacing some of the dense layers with equation-learning layers as detailed in [5]. As we are aware of many of the non-linear coupling functions that occur in the subtractable noise, this approach will hopefully make the networks easier to train and less prone to overfitting.

B.1 Bilinear Coupling

To illustrate, what a successfully trained network looks like, we start with a simple example, where the target channel y is the product of two auxillary channels x_1 and x_2:

$$y = x_1 x_2 \tag{6}$$

We give x_1 and x_2 white power spectra. We train the network with perfect witnesses, i.e. x_1 and x_2.

The results are shown in in Figure 7. Note that with just 1000 epochs of training, the subtractable noise is reduced by almost a factor of 10. From the downward trend of the loss plot, it is clear that we could do even better by training for longer and optimizing the learning rate.

B.2 Realistic angle-to-length noise

The network does not currently perform nearly as well for more realistic models. Here we show results for a realistic model of the angle-to-length noise. The realistic model includes 28 imperfect witness for the 8 true spot motion variables (pit and yaw on 4 interferometer mirrors) and the 4 true angular motion variables that strongly couple to the DARM. The true motions couple into the darm as essentially a sum of bilinaer couplings. The true motion and the witnesses are modeled with realistic power spectra that have a large dynamic range (as large as 10^5 for some features) with more power at low frequencies. To deal with the large dynamic range, we whiten (i.e. filter to reduce the dynamic range) the target and witness channels before training our network.

The results are shown in in Figure 8. From the ASD, we see that we only see that improvements are limited to less than a factor of 2 between 10 and 20 Hz. From the loss plot, we see that validation loss has plateaued and training for more epochs will likely not yield further improvement.

B.3 Obstacles

We suspect that the main reason why our network is performing worse with the realistic mock data is that the dynamic range of the features is very large. To illustrate this we return to the simple bilinear example from Appendix B.1. However instead of giving x_1 and x_2 white spectra, we give them red $1/f^2$ power spectra.

The results are shown in Figure 9. No real learning happens and the network overfits the training data (since the validation loss doesn't decrease). The only difference between this

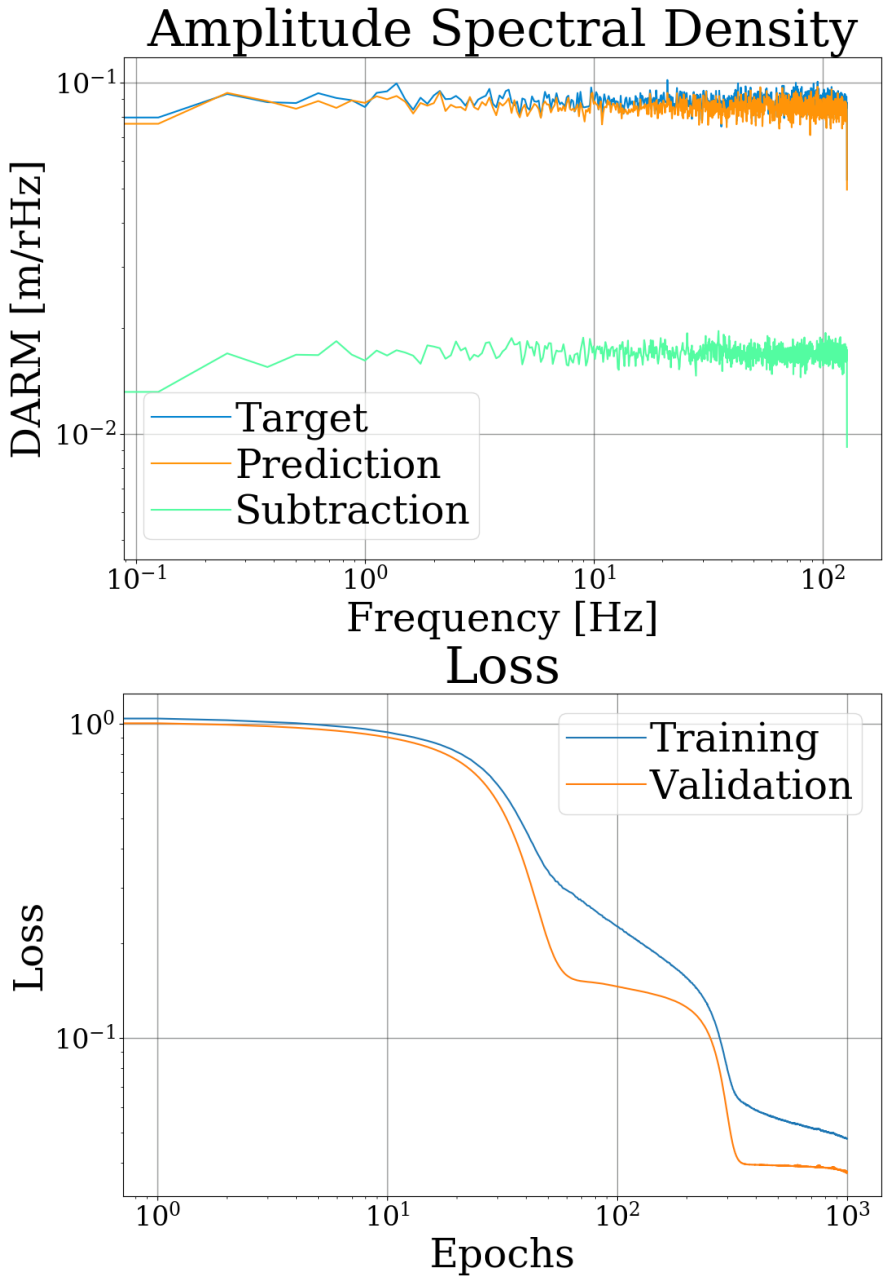


FIGURE 7: Results for simple bilinear coupling mock data (with features that have white PSD's). Top panel: Amplitude spectral density (for the validation data) for target (blue trace labeled "Data"), prediction from the network (yellow trace labeled "Pred"), and difference between the target and prediction (green trace labeled "Diff"). Bottom panel: Loss plot displaying training loss (blue trace) and validation loss (yellow trace). Note that one generally expects the validation loss to be above the training loss. However, when dropout layers are included Keras computes the training loss with the networks that are missing neurons and it computes the validation loss with the full network, allowing the validation loss to be lower than the training loss.

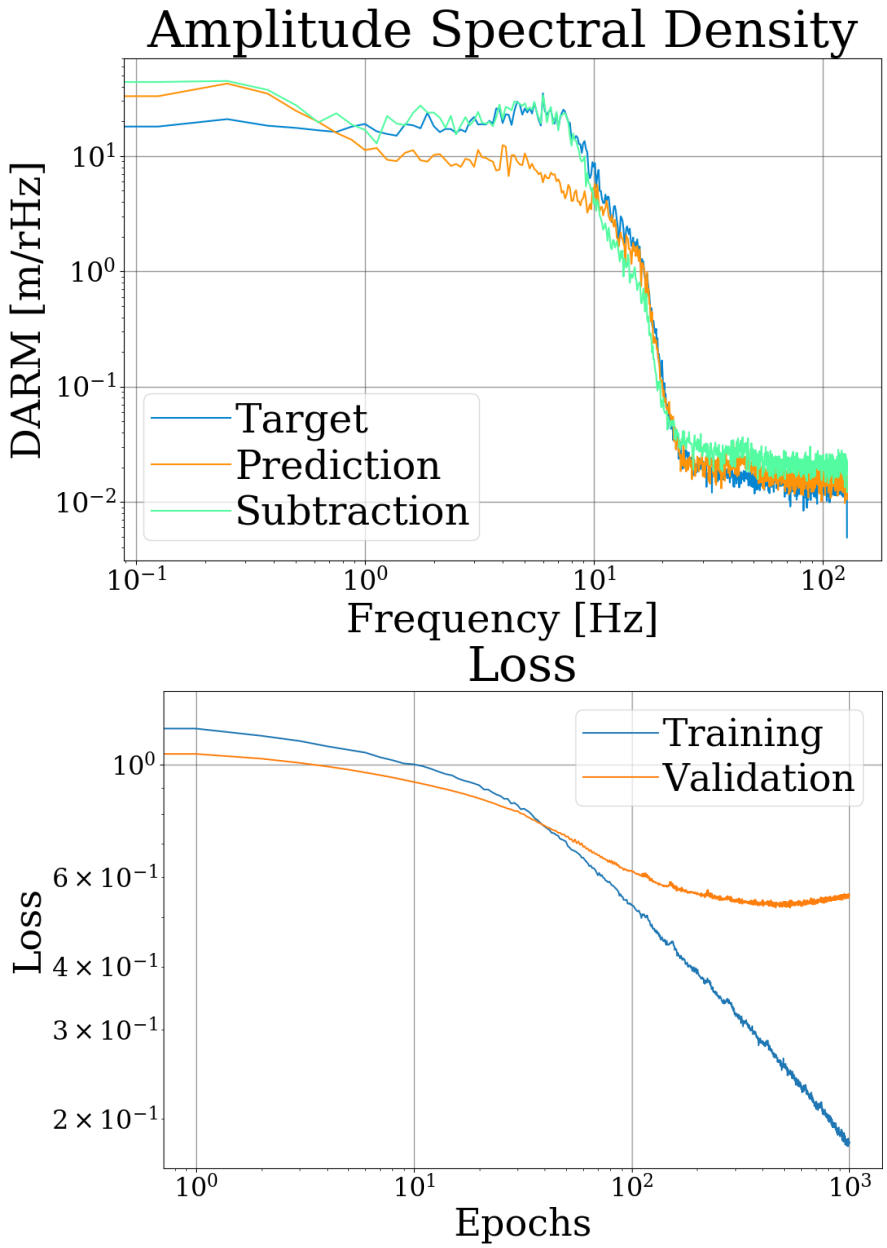


FIGURE 8: Results for realistic angle-to-length mock data. Top panel: Amplitude spectral density (for the validation data) for target (blue trace labeled "Data"), prediction from the network (yellow trace labeled "Pred"), and difference between the target and prediction (green trace labeled "Diff"). Bottom panel: Loss plot displaying training loss (blue trace) and validation loss (yellow trace).

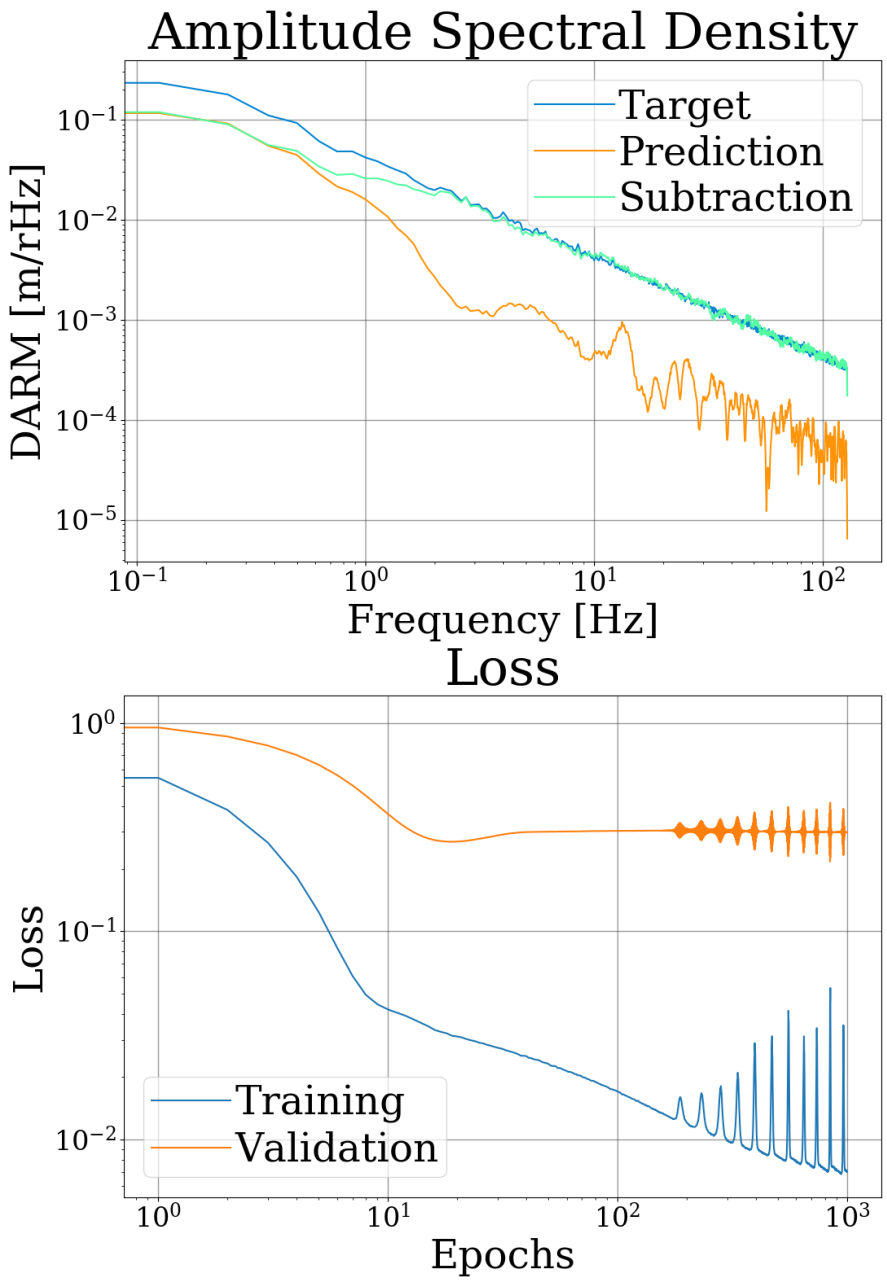


FIGURE 9: Results for simple bilinear coupling mock data with features that have a PSD $\propto 1/f^2$. Top panel: Amplitude spectral density (for the validation data) for target (blue trace labeled "Data"), prediction from the network (yellow trace labeled "Pred"), and difference between the target and prediction (green trace labeled "Diff"). Bottom panel: Loss plot displaying training loss (blue trace) and validation loss (yellow trace)

mock data from the mock data of Appendix B.1 is the power spectra of the input channels, so it is clear that is a problem.

We are trying to solve this problem with a combination of whitening the witness channels before input into the network and an intelligent choice of loss function. The high frequency portions of the input channels which have little power contribute to the target. By whitening the the witness channels before input into the network we can emphasize the high frequencies so that the network accounts for them during training. Similarly, by choosing the loss function intelligently, we can weight the frequencies that are important for gravitational wave detection, rather than the frequencies that contribute most to standard lost functions such as mean-squared-error.

References

[1] D. V. Martynov, E. D. Hall, B. P. Abbott, R. Abbott, T. D. Abbott, C. Adams, R. X. Adhikari, R. A. Anderson, S. B. Anderson, K. Arai, and et al. Sensitivity of the Advanced LIGO detectors at the beginning of gravitational wave astronomy. *PRD*, 93(11):112004, June 2016.

[2] Hang Yu. *Astrophysical signatures of neutron stars in compact binaries and experimental improvements on gravitational-wave detectors*. PhD thesis, Massachusetts Institute of Technology, 2019.

[3] Denis V. Martynov. *Lock Acquisition and Sensitivity Analysis of Advanced LIGO Interferometers*. PhD thesis, California Institute of Technology, 2015.

[4] François Chollet et al. Keras. https://keras.io, 2015.

[5] Samuel Kim, Peter Lu, Srijon Mukherjee, Michael Gilbert, Li Jing, Vladimir Ceperic, and Marin Soljacic. Integration of neural network-based symbolic regression in deep learning for scientific discovery, 2019.