The Neural LASSO: Local Linear Sparsity for Interpretable Explanations

Andrew Slavin Ross*
Harvard University
Cambridge, MA 02138
`andrew_ross@g.harvard.edu`

Isaac Lage*
Harvard University
Cambridge, MA 02138
`erikalage@g.harvard.edu`

Finale Doshi-Velez
Harvard University
Cambridge, MA 02138
`finale@seas.harvard.edu`

Abstract

Neural networks often perform better on prediction problems than simpler classes of models, but their behavior is difficult to explain. This makes it challenging to trust their predictions in safety critical domains. Recent work has focused on explaining their predictions using local linear approximations [1, 10], but these explanations can be complex when they depend on many features and it is unclear if they can be used to understand global trends in model behavior. In this work, we train neural networks to have sparse local explanations by applying L1 penalties to their input gradients. We show explanations of these networks depend on fewer inputs while their performance remains comparable across datasets and architectures. We illustrate how our approach encourages a different kind of sparsity than L1 weight decay. In a case study with ICU data, we observe that gradients vary smoothly over the input space, which suggests they can be used to gain insight into the global behavior of the model.

1 Introduction

Neural networks are the state of the art for many classification tasks. They work well for prediction problems with large datasets that depend on feature interactions and nonlinearities. But their expressivity comes at the cost of vulnerability to overfitting [13]. Held-out evaluations are effective but do not catch overfitting to biases shared between train and test sets. Without explanations to help domain experts identify these biases, neural networks may be unsafe for use in high risk domains [2].

An approach to interpreting the behavior of neural networks is to approximate the decision boundary with a set of interpretable models at many points throughout the input space [10]. However these explanations only capture very local trends in the model's behavior. Whether they can be used to gain a higher level understanding of how the model makes decisions is an open question.

We propose a penalty on neural networks that encourages the input gradients at each training data point to be sparse. We find that this makes input gradient based explanations more concise, and the variation of gradients across training points more structured, at no cost to accuracy. Section 2 outlines related work, Section 3 describes the model, and Section 4 is a cross-dataset comparison of our model with several others.

*Equal contribution.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Related Work

Interpreting neural networks is an active area of research. The approach that we build on generates explanations from local linear approximations of the decision boundary. **(author?)** [10] perturb each data point and fit a linear model to mimic the neural network's predictions. **(author?)** [1] interpret the predictions of a neural network by examining the input gradients of each feature. We build on **(author?)** [11] but without the need for expert annotation. Our work does not propose a new explanation method but rather a modification to the underlying network to encourage more coherent explanations when these methods are applied. Specifically, we use the same explanations as **(author?)** [1] but train with a penalty that encourages them to be sparse.

Adding an L1 penalty to the weights of logistic regression is a common way to learn sparse models [12]. These models depend on fewer features, which makes them more interpretable to humans [8]. The traditional neural network analogue applies a sparsity penalty to the weights of each neuron in the network–this is called weight decay. This penalty can have several effects: reducing overfitting, making learning easier, and making individual neurons more interpretable [3]. To the best of our knowledge, the effects of this penalty on the local linear approximations of neural networks have not been studied. We compare the effects of weight decay on input gradient explanations with our proposed penalty that directly encourages them to be sparse.

3 Our Method: Gradient LASSO

Neural networks learn a function $f(x|\theta)$ that makes predictions $\hat{y} \in \mathbb{R}^K$ given features $x \in \mathbb{R}^D$ about true labels $y \in \mathbb{R}^K$. We train them by searching for parameters θ that minimize a loss function usually defined as the cross entropy between our labels and predictions, $H(y, \hat{y})$.

In this paper, we add an additional term to the loss to encourage sparse local linear approximations. This takes the form of an L1 penalty to the gradients with respect to the sum of log probabilities across classes. This is proportional to the model's cross-entropy with a uniformly random guess, which can be interpreted as the model's level of certainty about its prediction. The gradient of this quantity, $-\nabla_x \sum_{k=1}^{K} \log f(x)_k$, also represents the score function with respect to its inputs–a classic measure of sensitivity. Empirically, we find that regularizing the gradient of the score function performs better than regularizing gradients of probabilities or log-odds. Our full loss function is:

$$\mathcal{L}(\theta|x, y) = H(y, \hat{y}) + \lambda_\theta \big|\big|\theta\big|\big|_1 + \lambda_\nabla \big|\big|\nabla_x H(\tfrac{1}{K}, \hat{y})\big|\big|_1, \tag{1}$$

where λ_θ controls the strength of the L1 penalty on our parameters and λ_∇ controls the strength of the L1 penalty on our explanation. In our experiments, at most one of these will be nonzero in a given model. We train by minimizing the average value of the loss across batches.

4 Empirical Evaluation

To study the effects of our regularization technique on neural network explanations, we conduct experiments on several datasets. The following are classic machine learning examples. With the Adult Census Income dataset [7], we predict whether yearly income is above $50,000 using z-scored census data. With the 20 Newsgroups Subset dataset [7], we predict whether an article is from the `alt.atheism` or `soc.religion.christian` newsgroup. We generate features by removing headers, footers, and quotes, and vectorizing examples using 5,000 dimensional one-hot word encodings selected with TF-IDF. With the MNIST [6] and CIFAR-10 [5] datasets, we predict the digit in an image using raw pixels as features.

We also conduct experiments with a synthetic dataset designed to demonstrate the capabilities of our method and a sepsis mortality prediction dataset where we conduct a more in-depth case study of how input gradients change across the input space. We construct the synthetic dataset to test that our model recovers the true explanation for data where labels depend on a small number of features that vary across the input space. We draw 49 dimensions from $\mathcal{N}(0, 1)$, and offset one of 6 region indicator dimensions by 10. The label is the sign of the product of 2 features determined by the region. We generate an equal number of samples from each region. For the sepsis task, we predict in-hospital mortality for patients from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) database [4]. We use demographics and 4-hour time slices of ICU readings selected and preprocessed according to the procedure in **(author?)** [9]. We robustly standardize the features to mitigate the influence of outliers and balance class labels, holding out test data at the patient level.

In Table 1, we report holdout AUC, network weight L1 norms, and D_{eff}, a measure of the number of relevant features. See the supplementary material for details about model architectures and training.

Dataset	D	Model	AUC	$\|\|\theta\|\|_1$	D_{eff}
Sparse Synthetic	49	Linear	0.498	0.53	32
		Linear, $\lambda_\theta = 100$	0.498	0.03	5
		MLP, Normal	0.943	734	8
		MLP, $\lambda_\theta = 0.0005$	0.961	279	5
		MLP, $\lambda_\nabla = 1:2$	0.994	386	2
Adult Income	89	Linear	0.904	11	13
		Linear, $\lambda_\theta = 100$	0.905	7	10
		MLP, Normal	0.905	792	35
		MLP, $\lambda_\theta = 0.0005$	0.910	55	15
		MLP, $\lambda_\nabla = 1:10$	0.910	406	9
Sepsis Mortality	47	Linear	0.809	4	29
		Linear, $\lambda_\theta = 100$	0.807	3	23
		MLP, Normal	0.708	740	35
		MLP, $\lambda_\theta = 0.0025$	0.816	32	27
		MLP, $\lambda_\nabla = 3:4$	0.827	360	20
Newsgroups Subset	5000	Linear	0.891	14593	1789
		Linear, $\lambda_\theta = 1$	0.798	128	38
		MLP, Normal	0.900	9611	4167
		MLP, $\lambda_\theta = 0.0005$	0.862	241	269
		MLP, $\lambda_\nabla = 1:100$	0.820	1488	35
			Accuracy		
MNIST	728	CNN_6, Normal	99.3%	80534	408
		CNN_6, $\lambda_\nabla = 1:10$	99.1%	71288	166
CIFAR-10	3072	CNN_9, Normal	80.9%	51337	1472
		CNN_9, $\lambda_\nabla = 1:10$	79.3%	49385	1432

Table 1: Cross-dataset comparison of heldout performance for different model types. We define the "effective number of features" D_{eff} as the average number of features whose input gradient magnitudes are at least $\frac{1}{10}$ of the largest for each example. Gradient-regularized networks match the sparsity of linear LASSO while maintaining predictiveness comparable to normal NNs.

Gradient LASSO achieves similar or greater predictive performance across a range of datasets and model architectures. Gradient regularization improves accuracy on the sparse synthetic and sepsis datasets, and does not significantly hurt accuracy on Adult Income, MNIST, and CIFAR-10. On 20 Newsgroups, gradient LASSO performs much worse than the unregularized neural network and logistic regression, suggesting that sparsity is not a useful prior for this dataset. However, it does achieve the same sparsity as logistic LASSO while outperforming it in AUC. These results suggest that a sparsity penalty can often be added to neural networks without a cost to accuracy.

Gradient LASSO and L1 weight decay encourage different types of sparsity. Across all rows of Table, 1, L1 weight decay and gradient regularization both shrink parameters θ and reduce the effective number of features D_{eff}, but $||\theta||_1$ is consistently lower for weight-regularized models and D_{eff} is consistently lower for gradient-regularized models. On the synthetic dataset, gradient-regularized model gradients are smaller and more axis-aligned than weight-regularized models' (Figure 1). On the sepsis dataset, we found that weight decay tended to encourage sparsity in hidden unit activations rather than features. See the supplementary material for visualizations of explanations for all datasets.

Gradient LASSO explanations exhibit smooth, clinically sensible contextual variation across ICU predictions. Figure 2 shows the sepsis mortality input gradients of several clinically interesting labs plotted against their values. Gradients of the gradient regularized neural network vary as a function of the lab value. The gradients of the weight-regularized model stay relatively constant, while

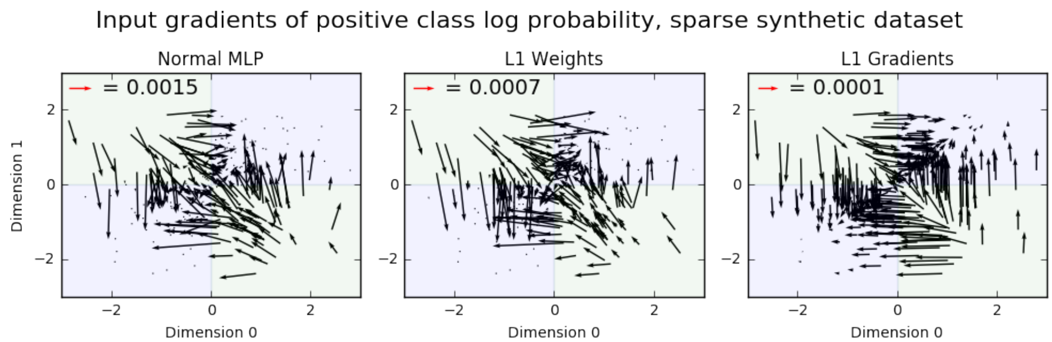

Figure 1: Positive class log probability input gradients for MLPs trained normally (left), with L1 weight regularization (middle), and with L1 gradient regularization (right) on sparse synthetic input examples in the first "region." Being in the 1st or 3rd quadrant determines class membership. Input gradients of gradient-regularized models are smallest and most axis-aligned, indicating simultaneous shrinkage and selection. They also exhibit the least variation with distance from the decision boundary.

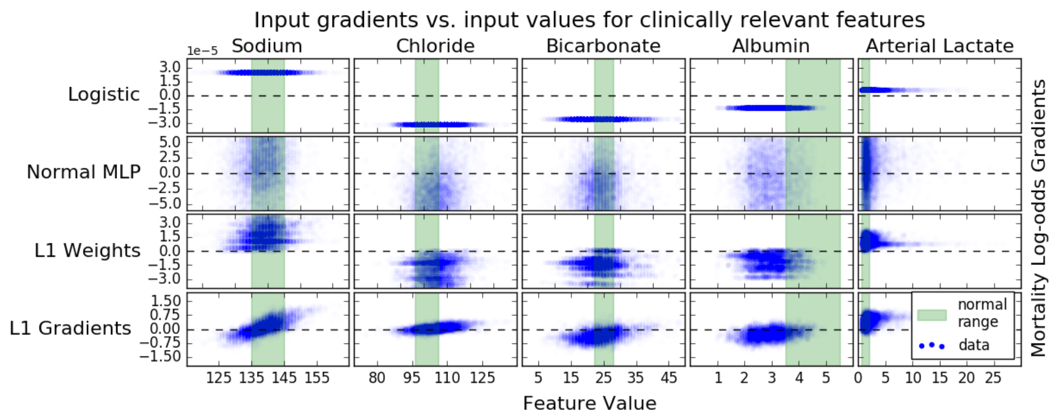

Figure 2: Sepsis test set lab values plotted against their examples' mortality gradients, with normal ranges for each feature overlaid. The gradient-regularized MLP often learns a smoothly-varying association between mortality risk and lab values outside their normal ranges. The weight-regularized MLP gradients are more discretized and similar to the logistic regression weights. They generally keeping the same sign regardless of feature value, even when clinically inappropriate.

the gradients of the unregularized model (which overfits) have less discernible structure. Although Figure 2 only captures variation in a single dimension, in the supplementary material we show that input gradient associations for gradient-regularized models exhibit smooth variations across the PCA projection of the input space. The smoothness of this variation suggests that local linear approximations can be used to gain insight into the global behavior of these networks. The variation itself suggests they are still flexible enough to capture important nonlinearities in the data.

Gradient regularized and weight regularized neural network explanations encourage different behaviors and may be useful in different cases. In future work, we plan to further explore their differences. In preliminary experiments, we also noticed that gradient-regularized neural networks are less certain in their predictions (see supplement). We believe this occurs because imposing a penalty on the gradient of the sum of log probabilities directly limits how quickly the model's certainty can change with changes in X, or alternatively because $\nabla_x \log f(x)_k = \frac{1}{f(x)_k}\nabla_x f(x)_k$ becomes too large when any predicted class probability $f(x)_k$ is too small. More balanced probabilities returned by the model do not necessarily affect accuracy, but they may affect how predictions are interpreted by end users. Regardless of whether this is desirable, our results demonstrate that we can regularize neural networks to be locally sparse without being globally sparse. This makes them easier to interpret without limiting their representational freedom.

Acknowledgments

IL was supported by NIH training grant 5T32LM012411-02.

References

[1] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

[2] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015.

[3] Geoffrey Hinton. A practical guide to training restricted boltzmann machines. *Momentum*, 9(1):926, 2010.

[4] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

[5] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[6] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The MNIST database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`, 2010.

[7] M. Lichman. UCI machine learning repository, 2013.

[8] Zachary Chase Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016.

[9] Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach. *arXiv:1705.08422*, 2017.

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[11] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2662–2670, 2017.

[12] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[13] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016.