Statistical correction of input gradients for black box models trained with categorical input features

Antonio Majdandzic [1] **Peter K. Koo** [1]

Abstract

Gradients of a model's prediction with respect to the inputs are used in a variety of downstream analyses for deep neural networks (DNNs). Examples include post hoc explanations with attribution methods. In many tasks, DNNs are trained on categorical input features subject to value constraints – a notable example is DNA sequences, where input values are subject to a probabilistic simplex constraint from the 1-hot encoded data. Here we observe that outside of this simplex, where no data points anchor the function during training, the learned function can exhibit erratic behaviors. Thus, the gradients can have arbitrary directions away from the data simplex, which manifests as noise in gradients. This can introduce significant errors to downstream applications that rely on input gradients, such as attribution maps. We introduce a simple correction for this off-simplex-derived noise and demonstrate its effectiveness quantitatively and qualitatively for DNNs trained on regulatory genomics data. We find that our correction consistently leads to a small, but significant improvement in gradient-based attribution scores, especially when the direction of the gradients deviates significantly from the simplex.

1. Introduction

Deep neural networks (DNNs) have been applied successfully to many regulatory genomics tasks, such as predicting the binding strength between proteins and DNA (Eraslan et al., 2019; Koo & Ploenzke, 2020). In general, two properties that are considered to be very important for DNNs are their predictive performance and interpretability, which is somewhat elusive but generally refers to the ability to explain the network's decision process. In practice, this is achieved by revealing simplified, human-interpretable representations that affect model predictions. In regulatory genomics, post hoc attribution methods are typically employed to provide importance scores for each nucleotide in a given sequence, often revealing motif-like representations that are important for model predictions (Shrikumar et al., 2017; Kelley et al., 2018; Nair et al., 2020; Avsec et al., 2021). The most popular and widely used methods often rely on gradients of the predictions with respect to the inputs, such as saliency maps (Simonyan et al., 2013) and integrated gradients (Sundararajan et al., 2017); these methods provide "importance scores" that represent the prediction sensitivity to each input feature.

Here we show that input gradients are prone to a specific type of noise when the input features have a geometric constraint set by a probabilistic interpretation, such as 1-hot-encoded DNA sequences. In such cases, all data lives on a lower-dimensional simplex within a higher-dimensional space; for DNA, the data lives on a 3D plane within a 4D space. A DNN has freedom to express any function shape off of the simplex, because no data points exist to guide the behavior of the function. This randomness can introduce unreliable gradient components in directions off the simplex, which can manifest as spurious noise in the input gradients, thereby affecting explanations from gradient-based attribution methods. We introduce a simple correction to minimize the impact of this off-simplex-derived gradient noise and show that in doing so, gradient-based attribution maps consistently improve both quantitatively and qualitatively.

2. Gradients for data that live on a simplex

Input features to DNNs in genomic prediction tasks are sequences represented as 1-hot encoded arrays of size $L \times 4$, having 4 nucleotide variants at each position of a sequence of length L (Fig. 1a). 1-hot encoded data naturally lends itself to a probabilistic interpretation, where each position corresponds to the probability of 4 nucleotides for DNA or 20 amino acids for proteins. While the values here represent definite/binary values, these 1-hot representations can also be relaxed to represent real numbers – this is a standard view for probabilisic modeling of biological sequences (Durbin et al., 1998), where the real numbers represent statistical

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory; Cold Spring Harbor, NY. Correspondence to: A Majdandzic <majdand@cshl.edu> and PK Koo <koo@cshl.edu>.

The 2021 ICML Workshop on Computational Biology.

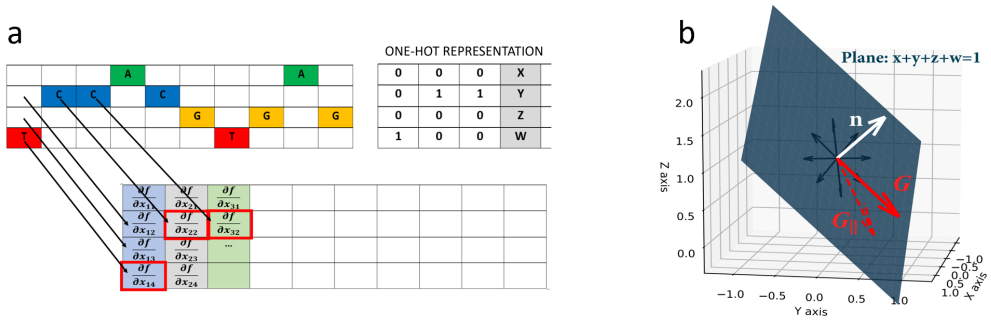

Figure 1. a) One-hot encoded genetic sequence example. General values (x, y, z, w) can be interpreted as probabilities. b) General geometric relation of the gradient and the simplex.

quantities like nucleotide frequencies. When the column at each nucleotide position is described by a vector of 4 real numbers – given by x, y, z , w – the probability axiom imposes that their sum is constrained to equal 1, that is

$$x + y + z + w = 1. \tag{1}$$

This restricts the data to a linear simplex (subspace) of allowed combinations, and Eq. 1 – being an equation of a 3D plane in a 4D space – defines this simplex. During training, a DNN is going to learn a function that is supported by the data that solely lives on this simplex, but it will have freedom to express any function shape outside of this plane, for which no training data exists. Since all data, including held-out test data, lives on this simplex, such a DNN can still maintain good predictions, despite its unregulated behavior off of the simplex. Nevertheless, when a function behaves erratic outside of the simplex, especially at points near the simplex where data lies, this could substantially affect input gradients. Thus, we hypothesize that off-simplex gradients introduce noise to attribution maps and other downstream applications that rely on input gradients.

The input gradients can be decomposed into two components: the component parallel to the data simplex (Eq. 1), which is supported by data, and the component orthogonal to the simplex, which we suspect is unreliable as the functions off the simplex are not supported by any data during training. We propose removing the unreliable orthogonal component from the gradient via a directional derivative, leaving only the parallel component that is supported by data. Without loss of generality, we now illustrate this procedure and derive the formula for this gradient correction in the case of widely used 1-hot encoded genomic data. Given $\vec{n} = \frac{1}{2}(\hat{i}+\hat{j}+\hat{k}+\hat{l})$ is a normal vector to the simplex plane (Eq. 1) and $\vec{G}$ is the gradient of function f,

$$\vec{G} = \frac{\partial f}{\partial x}\,\hat{i} + \frac{\partial f}{\partial y}\,\hat{j} + \frac{\partial f}{\partial z}\,\hat{k} + \frac{\partial f}{\partial w}\,\hat{l}\,, \tag{2}$$

we can correct $\vec{G}$ by removing the unreliable orthogonal component, according to:

$$\begin{aligned}\vec{G}_{\text{corrected}} &= \vec{G}_{||} = \vec{G} - \vec{G}_{\perp} = \vec{G} - (\vec{G} \bullet \vec{n})\vec{n} \\ &= (\frac{\partial f}{\partial x} - \mu)\hat{i} + (\frac{\partial f}{\partial y} - \mu)\hat{j} + (\frac{\partial f}{\partial z} - \mu)\hat{k} + (\frac{\partial f}{\partial w} - \mu)\hat{l}\end{aligned} \tag{3}$$

where $\mu = \frac{1}{4}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z} + \frac{\partial f}{\partial w}\right)$. By comparing Eqs. 2 and 3, we see that the corrected gradient at each position is obtained by simply subtracting the original gradient components by the mean gradients across components μ. Essentially, $\frac{\partial f}{\partial x}$ becomes $\frac{\partial f}{\partial x} - \mu$.

3. Experimental overview

To test whether our correction leads to more reliable attribution maps, we trained a DNN on a regulatory genomics task using synthetic data (Koo & Ploenzke, 2021), where we have ground truth. Specifically, the synthetic data reflects a simple billboard model of gene regulation (Slattery et al., 2014). Briefly, positive class sequences were embedded with 3 to 5 "core motifs" randomly selected with replacement from a pool of 5 known transcription factor motifs. Negative class sequences were generated in a similar way but with the exception that the pool of motifs also includes 100 non-overlapping "background motifs" from JASPAR (Mathelier et al., 2016). 20,000 sequences, each 200 nucleotides long, were randomly split into training, validation, and test sets according to 0.7, 0.1, and 0.2, respectively.

We used two different network architectures, namely CNN-shallow and CNN-deep from (Koo & Ploenzke, 2021), each with two variations – ReLU or exponential activations for the first convolutional layer – resulting in 4 models in total. CNN-shallow is a network that is designed to learn interpretable motifs in first layer filters with ReLU activations; while, CNN-deep is designed to learn distributed

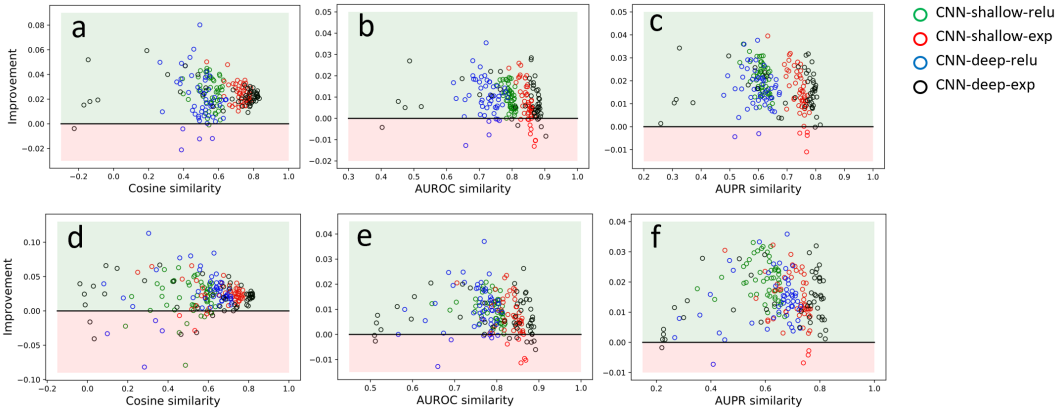

Figure 2. Interpretability performance. Improvement of similarity scores for saliency maps (a-c) and integrated gradient maps (d-f) for different similarity metrics. Improvement is defined as a change in similarity score after and before the correction. Light green region represents a positive improvement; light red is where the change in similarity score is worse. Scattered points represent 50 runs per each model: CNN-shallow-relu (green), CNN-shallow-exponential (red), CNN-deep-relu (blue), and CNN-deep-exponential (black).

motif representations (Koo & Eddy, 2019). Both networks learn robust motif representations in first layer filters when employing exponential activations. Models were trained in accordance with (Koo & Ploenzke, 2021).

We evaluate the efficacy of attribution maps by calculating similarity scores between the attribution maps and the ground truth. We investigated two different gradient-based attribution maps: saliency maps (Simonyan et al., 2013) and integrated gradients (Sundararajan et al., 2017), and applied several common similarity scores: cosine similarity, AUROC and AUPR. Cosine similarity uses a normalized dot product between the attribution map and the ground truth; the more similar the two maps are, the closer their cosine similarity is to 1. For AUROC and AUPR, we multiplied the attribution maps with the inputs, a so-called grad-times-input. For each sequence, we then generated a distribution of attribution scores at positions where ground truth motifs were embedded and a distribution of attribution scores at other positions. We quantified the separation of these two distributions using the AUROC and AUPR.

4. Results

By comparing the efficacy of attribution maps before and after correction for the 4 different CNN models (Sec. 3), we find that our gradient correction leads to a consistent improvement in interpretability, i.e. corrected attribution maps are consistently closer to the ground truth than the naive implementation. Figure 2a-c shows the consistent improvement of saliency maps across three similarity metrics for four different models each with 50 runs with different random initializations (light green regions represent positive improvement; light red regions represent a negative change in the similarity score). Figure 2d-f shows a similar improvement for integrated gradients. Evidently, our correction leads to consistent improvement in the attribution maps for each model.

To further support our hypothesis that noisy gradients arise from the angles between the naive input gradients and the simplex, we performed a statistical analysis of gradient angles with respect to the data simplex. Figure 3 summarizes our findings using CNN-shallow with exponential activations, and we obtain very similar results with other models. Figure 3a shows the probability density of gradient angles with respect to the simplex for positions that contain a ground truth motif; the distribution is centered around zero and the standard deviation is around 25 degrees. Most angles are small, therefore the trained function seems to produce gradients that naturally align close to the simplex. Figure 3b shows a scatter-plot of improvements vs angles for every position where a ground truth motif was embedded in 500 randomly chosen positive-label sequences – this was done for each model for all 50 runs. Here, improvement is given by the difference between the interpretability metric after correction minus before correction. Notice that the nucleotides that have large gradient angles with respect to the simplex are associated with a much larger improvement of attribution scores. The amount of correction is directly related to the angle: for 0 angle the correction is also 0, and this geometry results in the observed envelope where the true improvement (with respect to the ground truth) cannot exceed the amount of correction itself. We see that for most nucleotides with large angles, the improvement is near maximal (points are concentrated by the upper envelope with positive improvements). This also highlights how this correction only addresses off-simplex gradient noise. Addi-

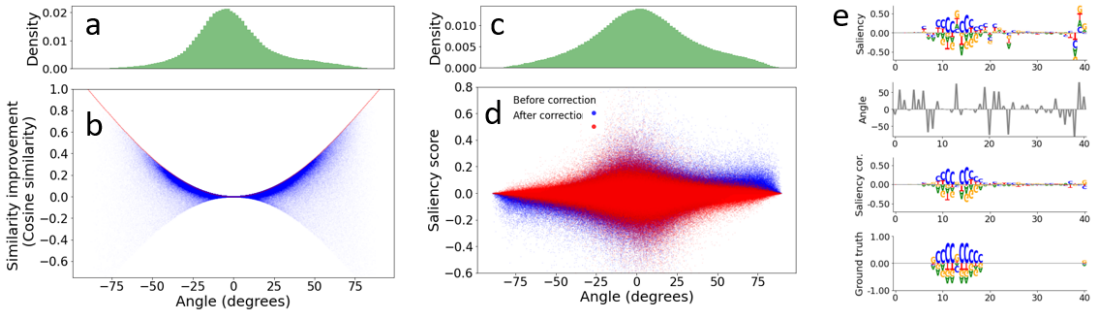

Figure 3. Analysis of gradients at different angles. (a, c) Probability density of input gradient angles for positions where ground truth motifs are embedded (a) and other background positions (c) across 500 randomly chosen positive-label test sequences. (b) Scatter plot of attribution score improvements based on cosine similarity (after correction minus before correction) versus the gradient angles for ground truth positions. Red line indicates the theoretical limit for a correction, i.e. $1 - \cos(\text{angle})$. (d) Scatter plot of saliency scores versus gradient angles before (blue) and after (red) correction for background positions (i.e. positions without any ground truth motifs). (b,d) Each dot represents a different nucleotide position. (e) Input gradient correction in action; from top to bottom: uncorrected saliency map, angles, corrected saliency map and ground truth.

tionally, since most angles are small and small angles lead to small corrections, this explains why our gradient correction method leads to a quite modest, though very consistent, improvement in attribution maps; most gradients are already closely aligned to the simplex.

Figure 3c shows the density of gradient angles with respect to the simplex for positions that do not contain a ground truth motif, i.e. background positions. Interestingly, the width of the background distribution of angles is broader than the distribution for ground truth positions. This suggests that background positions are more prone to off-simplex gradient noise, which creates spurious importance scores – a common feature observed in attribution maps for genomics. Figure 3d shows a scatter plot of the saliency score (i.e. grad-times-input) versus the angle for all background positions before correction (blue points) and after correction (red points) for the same sequences in Fig. 3b. Notice that after the correction, the saliency scores of nucleotides with large angles are greatly reduced, which we ascribe to down-weighting gradient noise in these positions. We also observe a large set of saliency scores near zero for which our correction method cannot address. We believe these represent false positive motifs that arise throughout this dataset simply by chance and so are not considered ground truth, despite exactly matching a ground truth motif pattern. To demonstrate how the gradient correction qualitatively affects attribution plots, in 3e we show a representative sequence patch from positive-label sequences. Uncorrected saliency maps (visualized as a sequence logo (Tareen & Kinney, 2020) – which shows positive and negative importance scores with a height that scales with the importance of that nucleotide) for CNN-deep-exp exhibits spurious noise throughout, especially at the positions directly flanking the ground truth motif pattern. After the correction, the spurious saliency scores in background positions, including the positions flanking the ground truth motifs, are driven towards zero, resulting in a (corrected) saliency map that better reflects the ground truth. Generally, improvements are visually discernable, thus the improvement is significant.

5. Discussion

Here we derived a simple gradient correction for data that live on a constrained simplex, and we demonstrate its effectiveness in gradient-based attribution methods. We find that our correction consistently leads to improvement in attribution scores. We emphasize that the noise removed is only the noise associated with erratic function behavior off-the-simplex. This correction is not a "magic bullet" that can correct other kinds of noise – i.e. if the function learns a noisy version of motifs for instance. The fact that the off-the-simplex gradient angles are typically small is itself a substantial and interesting property of the functions trained on categorical data with constraints.

Although our gradient correction formula was explicitly derived for the example of widely used 1-hot genomic data, our correction method – removing the components of the gradient orthogonal to the data simplex – is general and thus can be applied to any data structure with well defined geometric constraints, including protein sequences. Since our proposed correction is simple and can be incorporated in analysis pipelines with one line of code, we suggest all gradient-based methods should employ it, especially because it adds a little computational cost. Next, we plan to extend the study of our gradient correction to *in vivo* data where ground truth is not known.

References

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, 2021.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 2019.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.

Koo, P. K. and Eddy, S. R. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Computational Biology*, 15(12):e1007560, 2019.

Koo, P. K. and Ploenzke, M. Deep learning for inferring transcription factor binding sites. *Current Opinion in Systems Biology*, 2020.

Koo, P. K. and Ploenzke, M. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nature Machine Intelligence*, 3 (3):258–266, 2021.

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. Jaspar 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–D115, 2016.

Nair, S., Shrikumar, A., and Kundaje, A. fastism: Performant in-silico saturation mutagenesis for convolutional neural networks. *bioRxiv*, 2020.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2013.

Slattery, M., Zhou, T., Yang, L., Machado, A. C. D., Gordân, R., and Rohs, R. Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences*, 39(9):381–399, 2014.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Tareen, A. and Kinney, J. B. Logomaker: beautiful sequence logos in python. *Bioinformatics*, 36(7):2272–2274, 2020.