Saturating Auto-Encoder

Rostislav Goroshin
Courant Institute of Mathematical Science
New York University
goroshin@cs.nyu.edu

Yann LeCun
Courant Institute of Mathematical Science
New York University
yann@cs.nyu.edu

Abstract

We introduce a simple new regularizer for auto-encoders whose hidden-unit activation functions contain at least one zero-gradient (saturated) region. This regularizer explicitly encourages activations in the saturated region(s) of the corresponding activation function. We call these Saturating Auto-Encoders (SATAE). We show that the saturation regularizer explicitly limits the SATAE's ability to reconstruct inputs which are not near the data manifold. Furthermore, we show that a wide variety of features can be learned when different activation functions are used. Finally, connections are established with the Contractive and Sparse Auto-Encoders.

1 Introduction

An auto-encoder is a conceptually simple neural network used for obtaining useful data representations through unsupervised training. It is composed of an encoder which outputs a hidden (or latent) representation and a decoder which attempts to reconstruct the input using the hidden representation as its input. Training consists of minimizing a reconstruction cost such as L_2 error. However this cost is merely a proxy for the true objective: to obtain a useful latent representation. Auto-encoders can implement many dimensionality reduction techniques such as PCA and Sparse Coding (SC) [5] [6] [7]. This makes the study of auto-encoders very appealing from a theoretical standpoint. In recent years, renewed interest in auto-encoders networks has mainly been due to their empirical success in unsupervised feature learning [1] [2] [3] [4].

With only its reconstruction cost, the standard auto-encoder does not typically learn any meaningful hidden representation of the data. Well known theoretical and experimental results show that a linear auto-encoder with trainable encoding and decoding matrices, W^e and W^d respectively, learns the identity function if W^e and W^d are full rank or over-complete. The linear auto-encoder learns the principle variance directions (PCA) if W^e and W^d are rank deficient [5]. It has been observed that other representations can be obtained by regularizing the latent representation. This approach is exemplified by the Contractive and Sparse Auto-Encoders [3] [1] [2]. Intuitively, an auto-encoder with limited capacity will focus its resources on reconstructing portions of the input space in which data samples occur most frequently. From an energy based perspective, auto-encoders achieve low reconstruction cost in portions of the input space with high data density. If the data occupies some low dimensional manifold in the higher dimensional input space then minimizing reconstruction error achieves low energy on this manifold. Useful latent state regularizers raise the energy of points that do not lie on the manifold, thus playing an analogous role to minimizing the partition function in maximum likelihood models. In this work we introduce a new type of regularizer that does this explicitly for auto-encoders with a non-linearity that contains at least one flat (zero gradient) region. We show examples where this regularizer and the choice of nonlinearity determine the feature set that is learned by the auto-encoder.

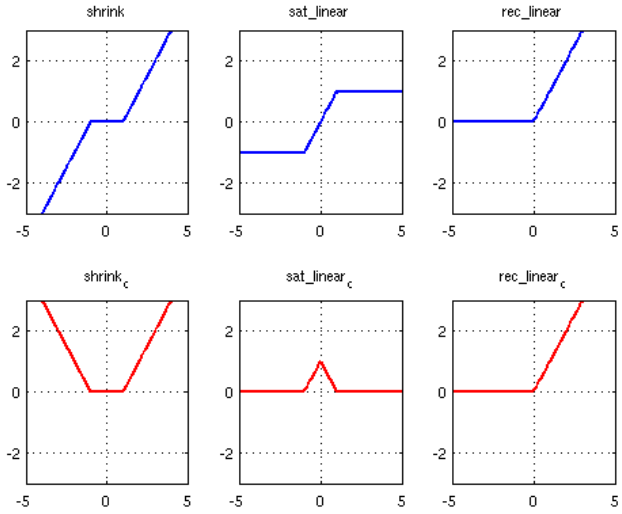


Figure 1: Three nonlinearities (top) with their associated complementary regularization functions(bottom).

2 Hidden Variable Regularization

Several auto-encoder variants which regularize their latent states have been proposed, they include the sparse auto-encoder and the contractive auto-encoder [1] [2] [3]. The sparse auto-encoder includes an over-complete basis in the encoder and imposes a sparsity inducing (usually L_1) penalty on the hidden activations. This penalty prevents the auto-encoder from learning to reconstruct all possible points in the input space and focuses the expressive power of the auto-encoder on representing the data-manifold. Similarly, the contractive auto-encoder avoids trivial solutions by introducing an auxiliary penalty which measures the square Frobenius norm of the Jacobian of the latent representation with respect to the inputs. This encourages a constant latent representation except around training samples where it is counteracted by the reconstruction term. It has been noted in [3] that these two approaches are strongly related. The contractive auto-encoder explicitly encourages small entries in the Jacobian, whereas the sparse auto-encoder is encouraged to produce mostly zero (sparse) activations which can be designed to correspond to mostly flat regions of the nonlinearity, thus also yielding small entries in the Jacobian.

2.1 Saturating Auto-Encoder through Complementary Nonlinearities

Our goal is to introduce a simple new regularizer which will explicitly raises reconstruction error for inputs not near the data manifold. Consider activation functions with at least one flat region; these include shrink, rectified linear, and saturated linear (Figure 1). Auto-encoders with such nonlinearities lose their ability to accurately reconstruct inputs which produce activations in the saturation regime(s) of their activation functions. With this in mind, we introduce a penalty of the form $f_c(\sum_{j=1}^{d} W_{ij}^e x_j + b_i)$ encourages the argument to be in the saturation regime of the activation function (f). We call this the Saturating Auto-Encoder (SATAE). For activation functions with zero-gradient regime(s) the complementary nonlinearity (f_c) can be defined as the distance to the nearest saturation region. Specifically, let $S = \{x \mid f'(x) = 0\}$ then we define $f_c(x)$ as:

$$f_c(x) = \inf_{y \in S} |x - y|. \tag{1}$$

Figure 1 shows three activation functions and their associated complementary nonlinearities. The complete loss to be minimized by a SATAE with nonlinearity f is:

$$L = \sum_{x \in D} \frac{1}{2} \|x - x_r\|^2 + \eta \sum_{i=1}^{d_h} f_c(W_i^e x + b_i^e), \tag{2}$$

where $x_r = W^d f(W^e x + b^e) + b^d$ is the reconstructed x for an auto-encoder with no output nonlinearity, and d_h denote the number of hidden units. The hyper-parameter η regulates the trade-off between reconstruction and saturation.

3 Effect of the Saturation Regularizer

We will examine the effect of the saturation regularizer on auto-encoders with a variety of activation functions. It will be shown that the choice of activation function is a significant factor in determining the type of basis the SATAE learns. First, we will present results on toy data in two dimensions followed by results on higher dimensional image data.

3.1 Visualizing the Energy Landscape

Given a trained auto-encoder the reconstruction error can be evaluated for a given input x. For low-dimensional spaces ($\mathbb{R}^n$, where $n \leq 3$) we can evaluate the reconstruction error on a regular grid in order to visualize the portions of the space which are well represented by the auto-encoder. More specifically we can compute $E(x) = \frac{1}{2}\|x - x_r\|^2$ for all x within some bounded region of the input space. Ideally, the reconstruction energy will be low for all x which are in training set and high elsewhere. Figures 2 and 3 depict the resulting reconstruction energy for inputs $x \in \mathbb{R}^2$, and $-1 \leq x_i \leq 1$. Black corresponds to low reconstruction energy. The training data consists of a one dimensional manifold shown overlain in yellow. Figure 2 shows a toy example for a SATAE which uses ten basis vectors and a shrink activation function. Note that adding the saturation regularizer decreases the volume of the space which is well reconstructed, however good reconstruction is maintained on or near the training data manifold. The auto-encoder in Figure 3 contains two encoding basis vectors (red), two decoding basis vectors (green), and uses a saturated-linear activation function. The encoding and decoding bases are unconstrained. The unregularized auto-encoder learns an orthogonal basis with a random orientation. The region of the space which is well reconstructed corresponds to the outer product of the linear regions of two activation functions; beyond that the error increases quadratically with the distance. Including the saturation regularizer however induces the auto-encoder to operate in the saturation regime at the extreme points of the training data, limiting the space which is well reconstructed. Note that because the encoding and decoding weights are separate and unrestricted, the encoding weights were scaled up to effectively reduce the width of the linear regime of the nonlinearity.

3.2 SATAE-shrink

Consider a SATAE with a shrink activation function and shrink parameter λ. The corresponding complementary nonlinearity, derived using Equation 1 is given by:

$$shrink_c(x) = \begin{cases} abs(x), & |x| > \lambda \\ 0, & \text{elsewhere} \end{cases} .$$

Note that $shrink_c(W^e x + b^e) = abs(shrink(W^e x + b^e))$, which corresponds to an L_1 penalty on the activations. Thus this SATAE is equivalent to a sparse auto-encoder with a shrink activation function. Given the equivalence to the sparse auto-encoder we anticipate the same scale ambiguity which occurs with L_1 regularization. This ambiguity can be avoided by normalizing the decoder weights to unit norm. It is expected that the SATAE-shrink will learn similar features to those obtain with a sparse auto-encoder, and indeed this is what we observe. Figure 4(a) shows 25 randomly selected decoder filters learned by an auto-encoder with shrink nonlinearity trained on natural 12x12 image patches. One can recognize the expected Gabor-like features when the saturation penalty is activated. When trained on the binary MNIST dataset the learned basis is comprised of portions of digits and strokes. Nearly identical results are obtained with a SATAE which uses a rectified-linear activation function. This is because a rectified-linear function with a bias can behave like as a

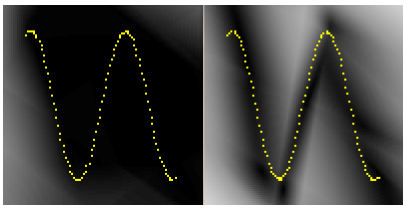


Figure 2: Energy surfaces for unregularized (left), and regularized (right) solutions obtained using the $shrink$ nonlinearity and 10 basis vectors. Black corresponds to low reconstruction energy. Training points lie on a one-dimensional manifold shown in yellow.

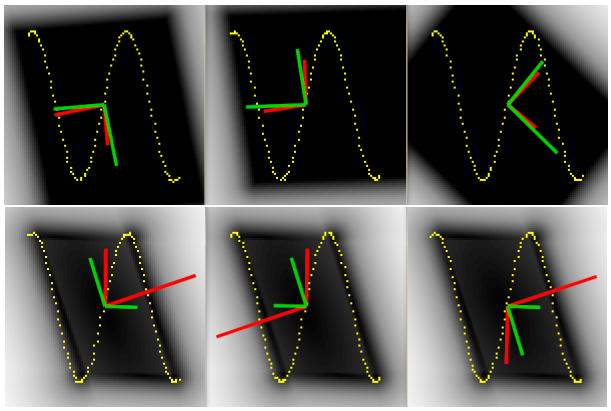
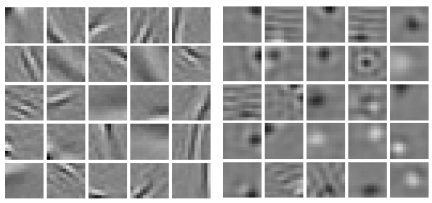


Figure 3: Top Row: three randomly initialized solutions obtained with no regularization. Bottom Row: three randomly initialized solutions obtained with regularization.

positive only shrink function, similarly the complementary function is equivalent to a positive only L_1 penalty on the activations.

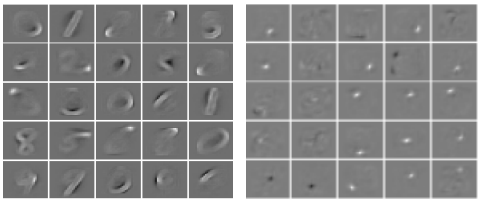
3.3 SATAE-saturated-linear

The SATAE with saturated-linear activation function learns a completely different feature set. Empirically, it was observed that increasing the regularization penalty produces more localized feature detectors. To understand why these feature detectors arise consider a dataset in which the variables take on binary values (e.g. MNIST). The scaled identity basis is a global minimizer of Equation 2 when a saturated-linear activation function is used. Such a basis can perfectly reconstruct any binary input while operating exclusively in the saturated regions of the activation function, thus incurring no saturation penalty. This is exactly the type of basis that arises in experiments when training on MNIST, see Figure 5b. Training on image patches produces a similar but less pronounced effect: more localized features are obtained but of greater variety (Figure 4b). In contrast to the SATAE-



(a) (b)

Figure 4: (a)Twenty-five randomly selected basis elements learned by the SATAE with shrink nonlinearity trained on 12x12 natural image patches. (b) Identical to (a), except that the SATAE uses a saturated-linear nonlinearity



(a) (b)

Figure 5: (a)Twenty-five randomly selected basis elements learned by the SATAE with shrink nonlinearity trained on 28x28 binary MNIST digits. (b) Identical to (a), except that the SATAE uses a saturated-linear nonlinearity

shrink, the SATAE-saturated-linear receives the heaviest penalty when the activation is zero, which tends to spread the responsibility of reconstructing the data among all the basis elements.

4 Experimental Details

In all experiments data samples were normalized by subtracting the mean and dividing by the standard deviation of the dataset. Experiments on MNIST were performed with 200 basis elements, while experiments on natural image patches used only 100 basis elements. The decoder basis elements of the SATAEs with shrink and rectified-linear nonlinearities were reprojected to the unit sphere after every 10 stochastic gradient updates. The SATAEs which used saturated-linear activation function were trained with tied weights. All results presented were obtained using stochastic gradient descent.

5 Discussion

We have demonstrated that by using different activation functions drastically different feature sets are learned by SATAEs. Hybrid SATAEs which use a mixture of activation functions also possible. The utility of these features depend on the application.

5.1 Relationship with the Contractive Auto-Encoder

Let h_i be the output of the i^{th} hidden unit of a single-layer auto-encoder with point-wise nonlinearity $f(\cdot)$. The regularizer imposed by the contractive auto-encoder (CAE) can be expressed as follows:

$$\sum_{ij} \left(\frac{\partial h_i}{\partial x_j} \right)^2 = \sum_{i}^{d_h} \left(f'(\sum_{j=1}^{d} W_{ij}^e x_j + b_i)^2 \|W_i^e\|^2 \right),$$

where x is a d-dimensional data vector, $f'(\cdot)$ is the derivative of $f(\cdot)$, b_i is the bias of the i^{th} encoding unit, and W_i^e denotes the i^{th} row of the encoding weight matrix. The first term in the above equation tries to adjust the weights so as to push the activations into the low gradient (saturation) regime of the nonlinearity, but is only defined for differentiable activation functions. Therefore the CAE indirectly encourages operation in the saturation regime. Computing the Jacobian, however, can be cumbersome for deep networks. Furthermore, the complexity of computing the Jacobian is $O(d \times d_h)$ [3], compared to the $O(d_h)$ for the saturation penalty.

5.2 Relationship with the Sparse Auto-Encoder

In Section 3.2 it was shown that SATAEs with shrink or rectified-linear activation functions are equivalent to a sparse auto-encoder. Like the sparsity penalty, the saturation penalty can be applied at any point in a deep network at the same computational cost. Unlike the sparsity penalty, the saturation penalty is adapted to the nonlinearity of the particular layer to which it is applied.

References

[1] Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra and Yann LeCun. Efficient Learning of Sparse Representations with an Energy- Based Model, in J. Platt et al. (Eds), *Advances in Neural Information Processing Systems (NIPS 2006)*, 19, MIT Press, 2006.

[2] Marc'Aurelio Ranzato, Fu-Jie Huang, Y-Lan Boureau and Yann LeCun: Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition, Proc. *Computer Vision and Pattern Recognition Conference (CVPR'07)*, IEEE Press, 2007

[3] Rifai, S. and Vincent, P. and Muller, X. and Glorot, X. and Bengio, Y. Contractive auto-encoders: Explicit invariance during feature extraction, *Proceedings of the Twenty-eight International Conference on Machine Learning, ICML 2011*

[4] P. Vincent, H. Larochelle, Y. Bengio, P.A. Manzagol. Extracting and Composing Robust Features with Denoising Autoencoders *Proceedings of the 25th International Conference on Machine Learning (ICML'2008)*, 2008.

[5] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, New York: John Wiley & Sons, 2001, pp. xx + 654, ISBN: 0-471-05669-3

[6] Olhausen, Bruno A.; Field, David J. (1997). Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?. *Vision Research 37 (23): 3311-3325.*

[7] Karol Gregor and Yann LeCun: Learning Fast Approximations of Sparse Coding, Proc. *International Conference on Machine learning (ICML'10)*, 2010