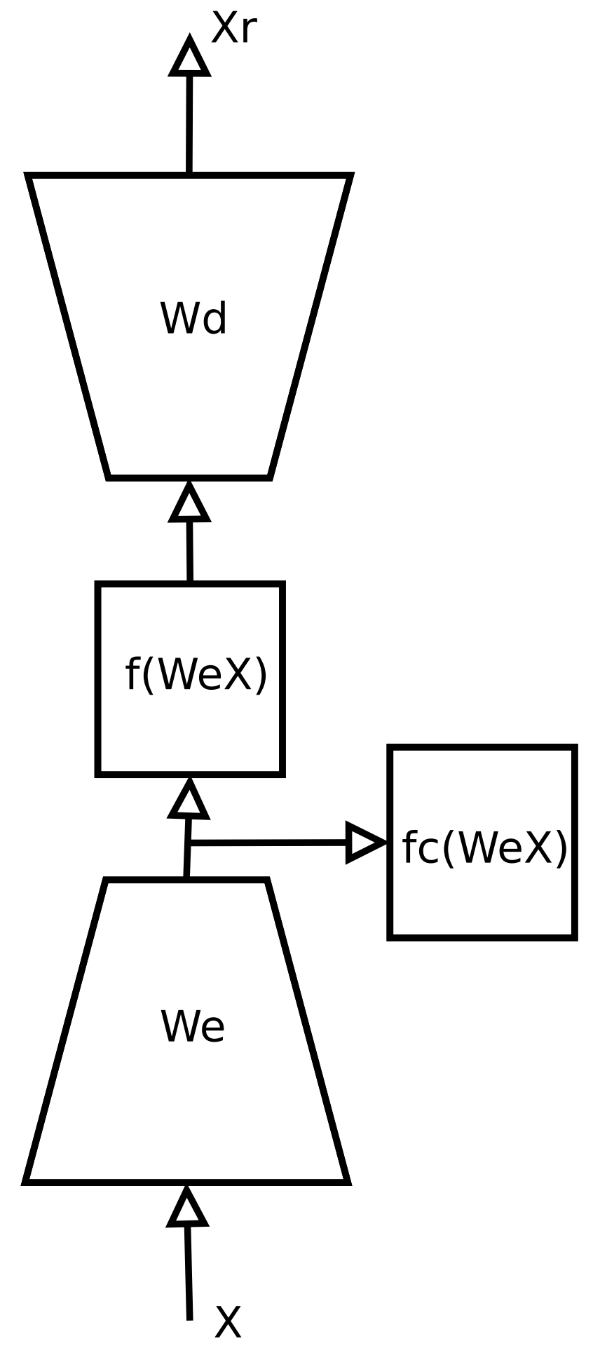
Saturating Auto-Encoders

Rostislav Goroshin and Yann LeCun
Courant Institute of Mathematical Sciences, New York University
{goroshin,yann}@cims.nyu.edu

Auto-Encoders Learn Manifolds

- Data lies on a low-dimensional manifold embedded in a higher dimensional ambient space
- Reconstruction objective ensures low reconstruction error in portions of the input space where data is densely distributed
- **Regularizers should raise reconstruction error for inputs not near the data manifold**
- Analogous to minimizing the partition function in max-likelihood models

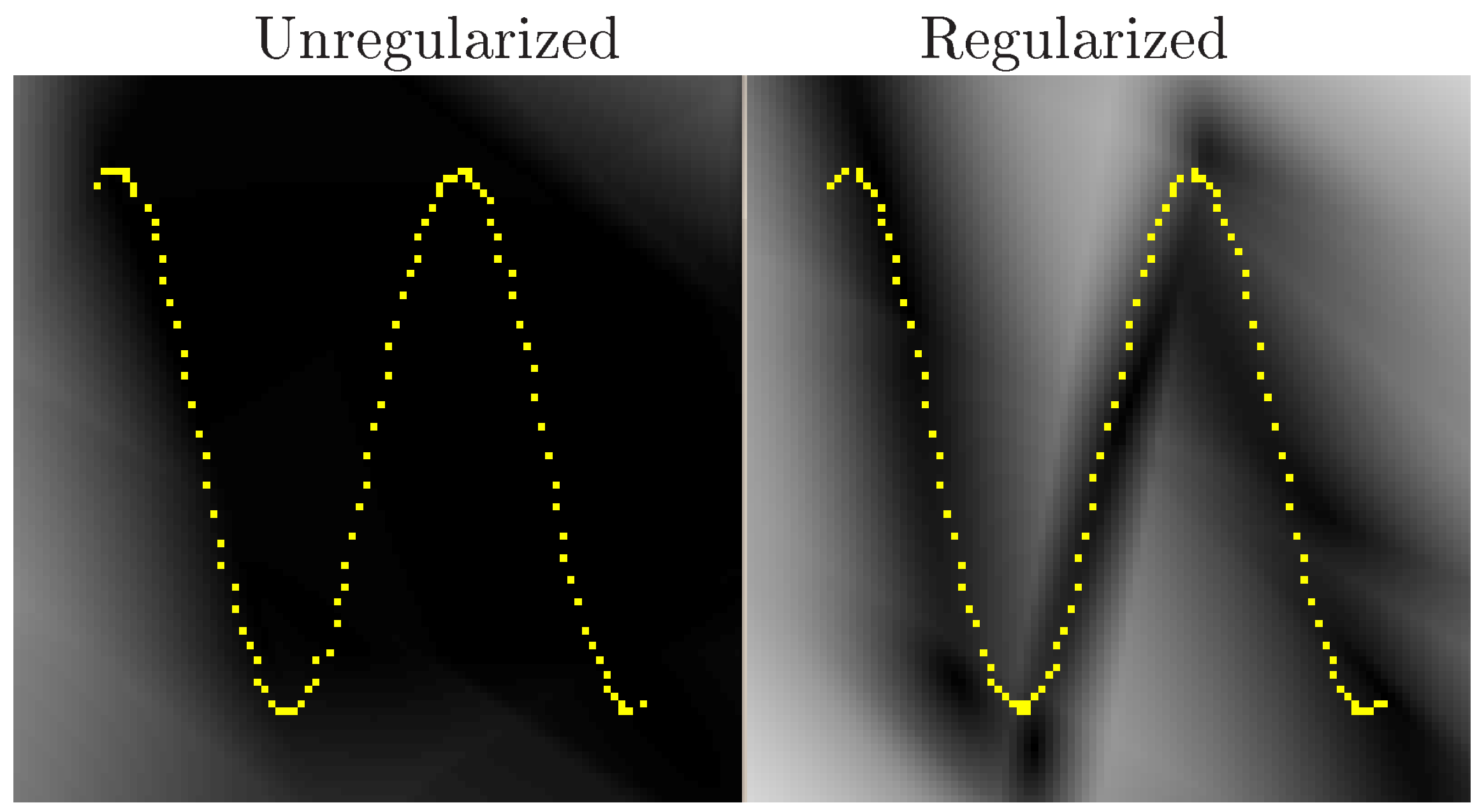


Latent State Regularization

Latent state regularization is a method of introducing an information bottleneck, which is more intuitive than regularizing the weights directly.

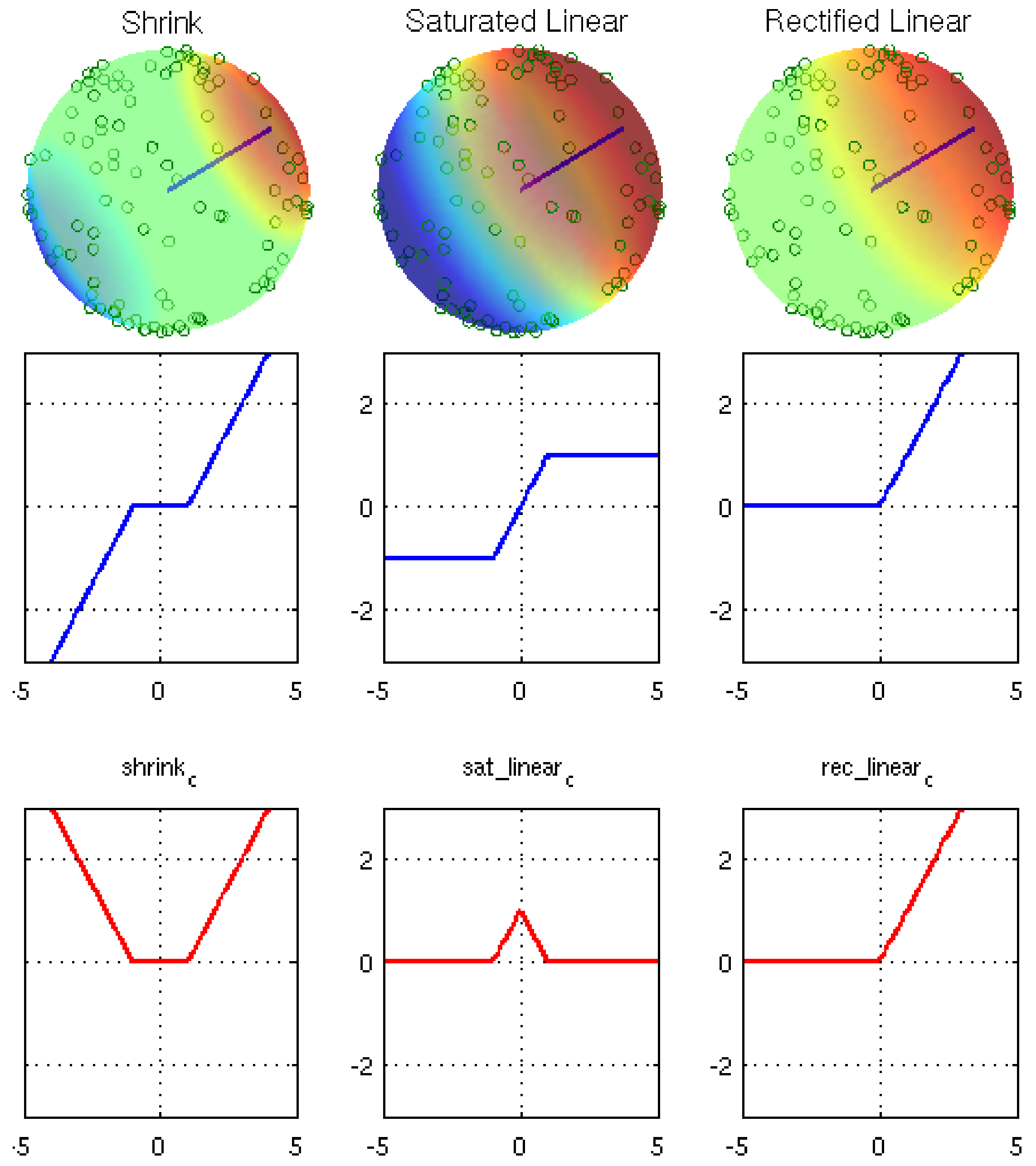
- Unsupervised Setting: Sparsity (L_1) and Contractive Regularization
- Supervised Setting: Penalize the Surface Area of the Decision Boundary

Toy-Manifold Example



Regularization via Saturation

- Consider activation functions with flat (zero-gradient) regions
- These activation functions lose their ability to reconstruct the input when activations occur in the flat regions
- Reconstruction error will increase quadratically as we move away from the manifold
- We associate with each activation function a complimentary function which encourages activations in the flat regions



Complimentary Function

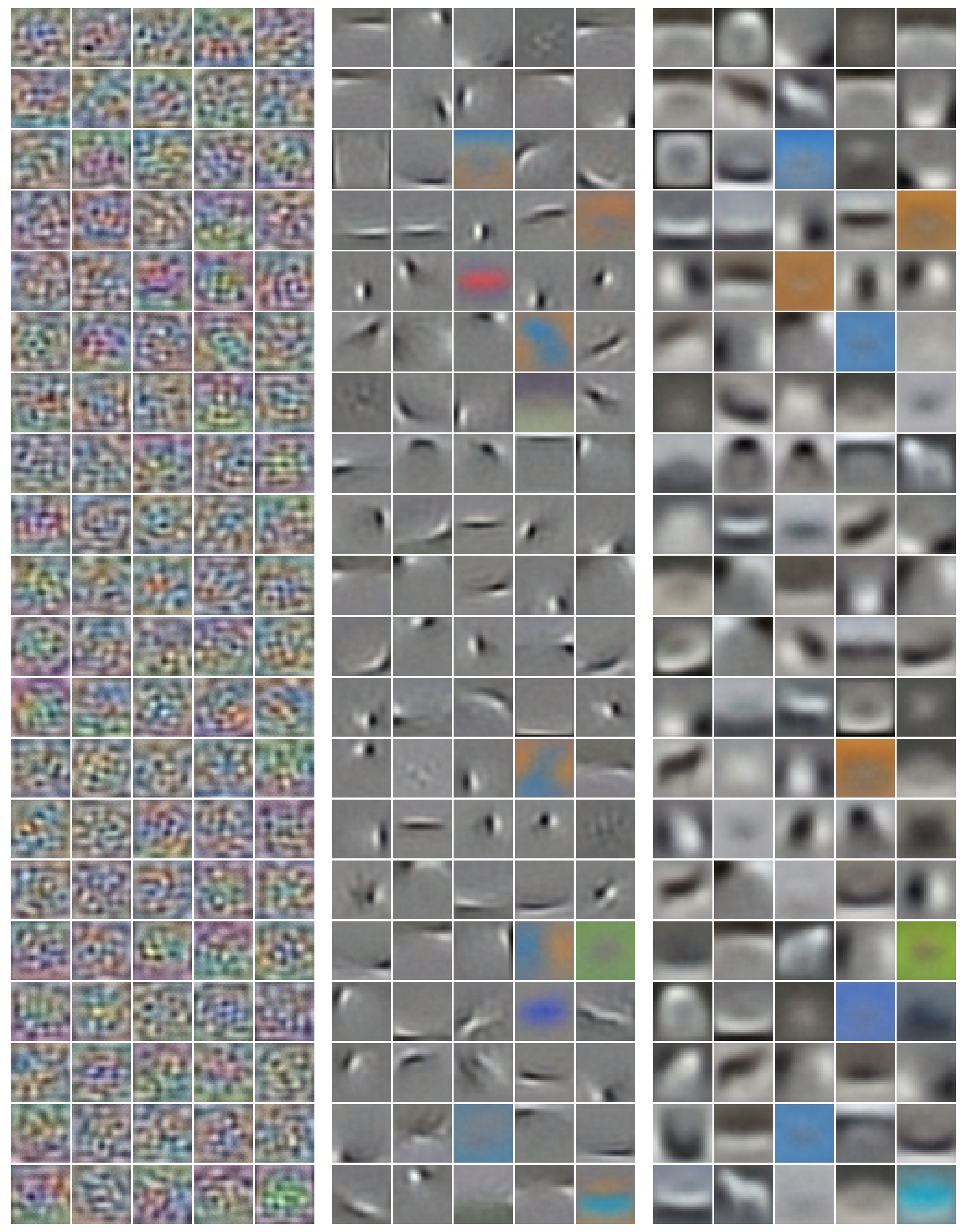
$$f_c(z) = \inf_{z' \in S} |z - z'|$$

Loss Functional

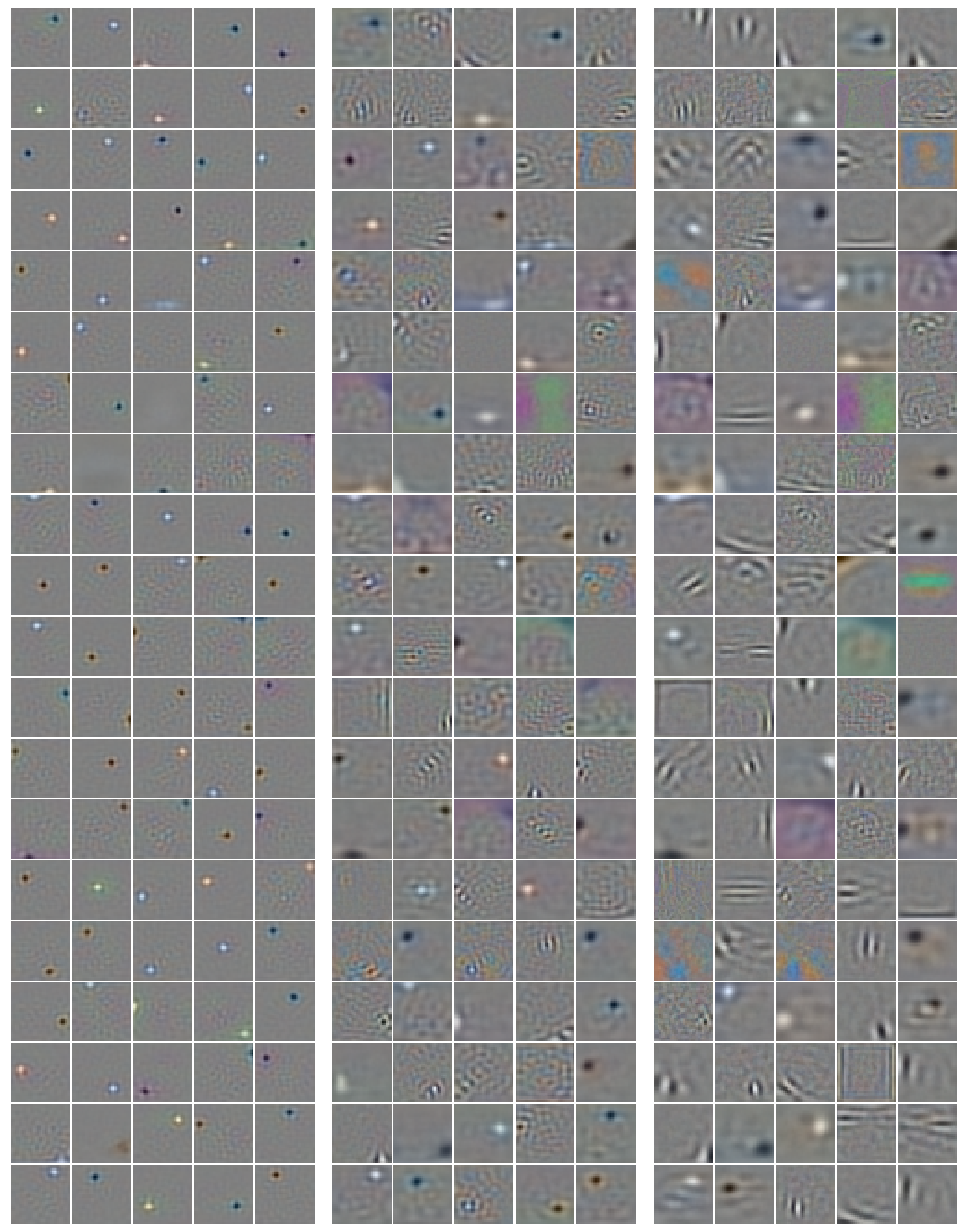
$$L = \sum_{x \in D} \frac{1}{2} \|x - \left(W^d f(W^e x + B^e) + B^d\right)\|^2$$

$$+ \ \alpha \sum_{i=1}^{d_h} f_c(W_i^e x + b_i^e)$$

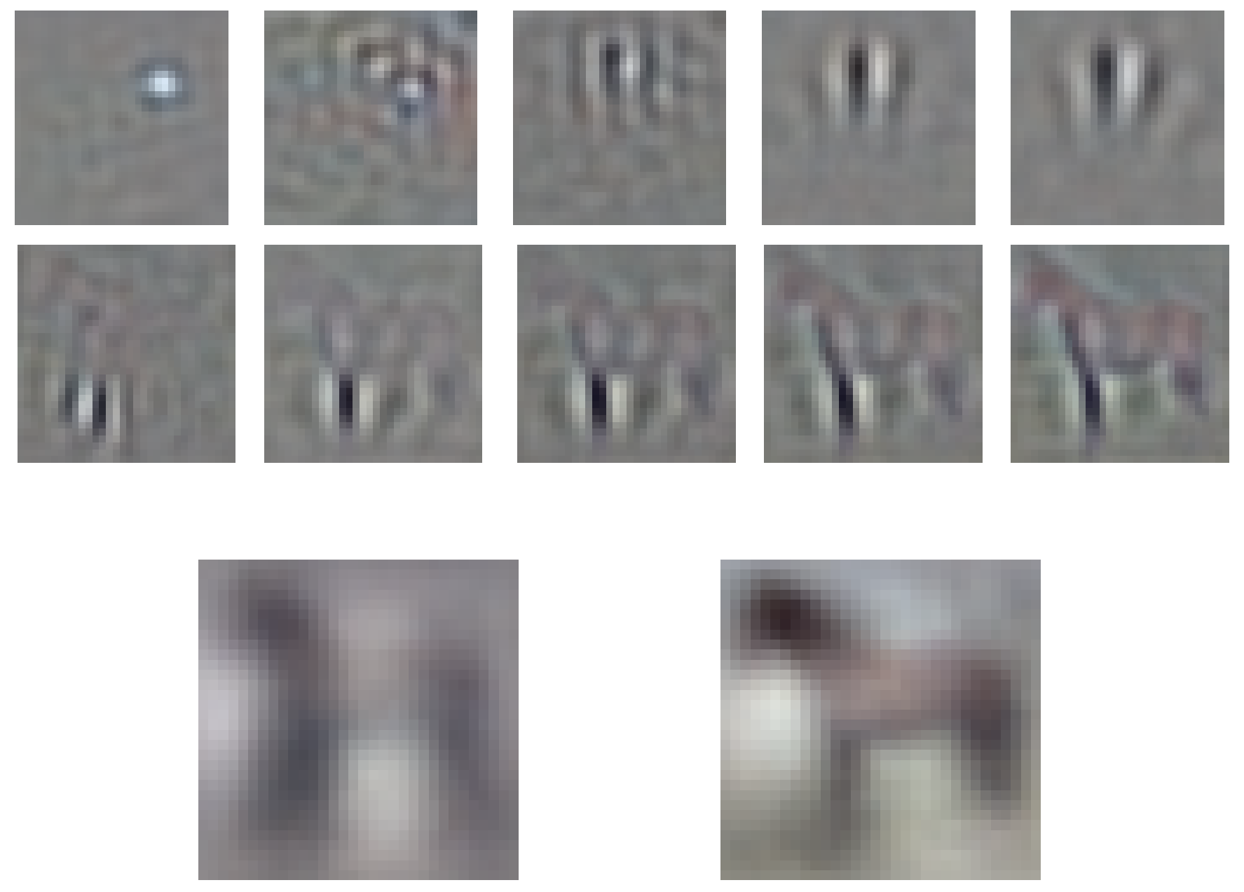
SATAE-shrink on CIFAR-10



SATAE-sat-linear on CIFAR-10



Progressively Increasing α



Relation to Other Regularizers

Connection to Sparse-Auto-Encoders

Note that $shrink_c(W^e x + b^e) = abs(shrink(W^e x + b^e))$. For *shrink*-nonlinearity, our regularizer corresponds to L_1 penalty on the activations.

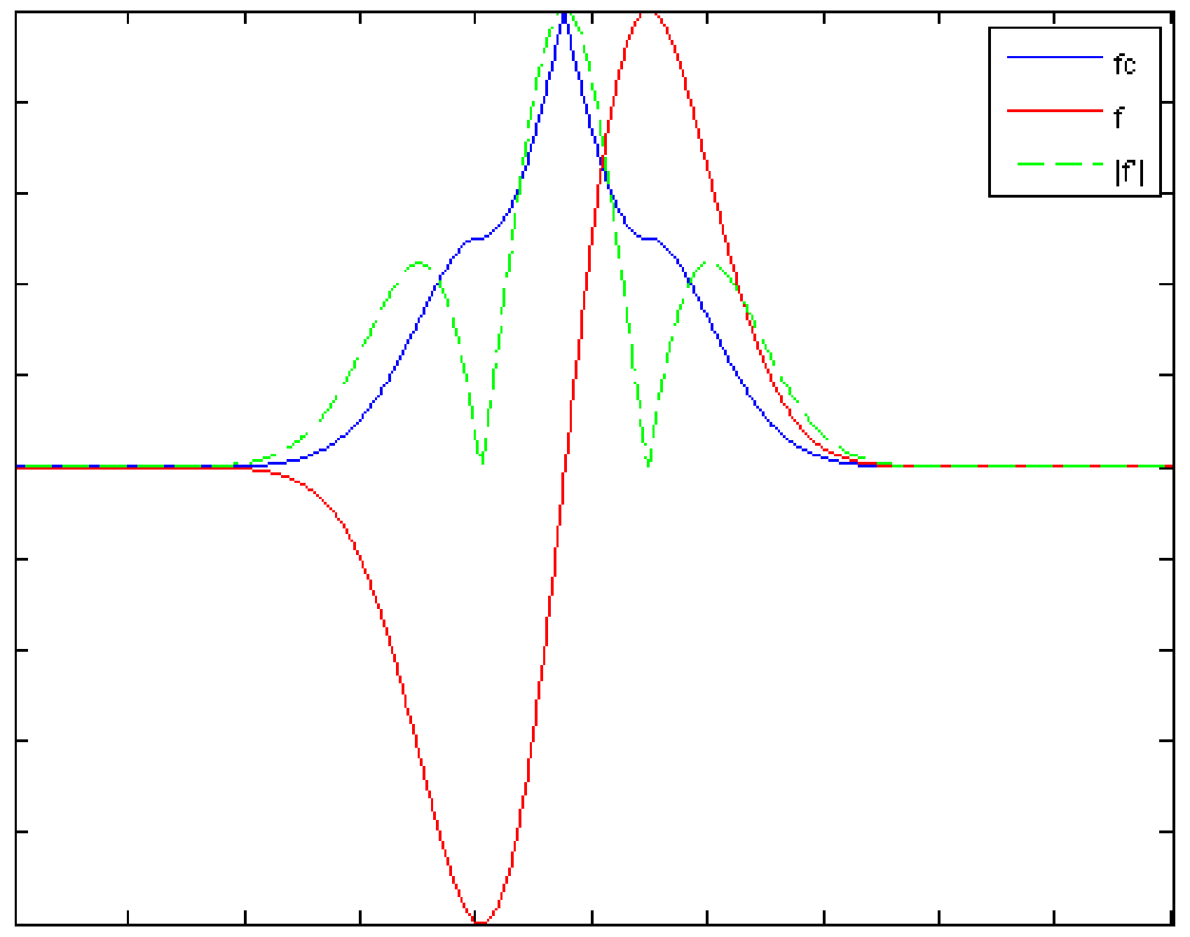
Connection to Contractive-Auto-Encoders

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

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.

Extension to C^1

We employ the concept of average variation over a finite interval to extend the definition of complimentary activation functions to differentiable functions.



References

1. 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*
2. Karol Gregor and Yann LeCun: Learning Fast Approximations of Sparse Coding, Proc. *International Conference on Machine learning (ICML'10)*, 2010
3. 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), *(NIPS 2006)*, 19, MIT Press, 2006.