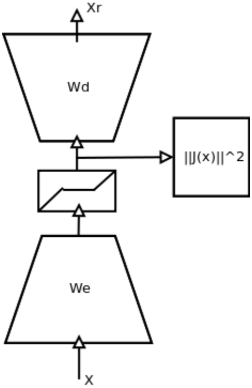
Contractive Auto-Encoder

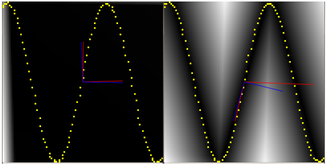
- A technique of obtaining "interesting" representations is to regularize the latent representation
- Contrary to more traditional regularization directly on weights, these correspond to "generic prior hypotheses"
- $L_{CAE} = \sum_{x \in D_n} \|x - x_r\|_2^2 + \lambda \sum_{ij} \left(\frac{\partial h_j(x)}{\partial x_i} \right)^2$



Contractive AE

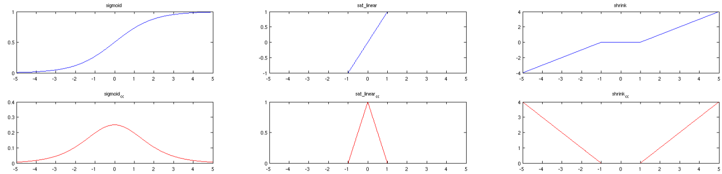
- $\left(\frac{\partial h_j(x)}{\partial x_i}\right)^2 = (h_j'(x))^2 W_{e_{ij}}^2$
- Important to tie or normalize weights
- If there were no reconstruction objective then the penalty on the Jacobian would produce a constant representation for all inputs either by saturating or weight decay
- However nearby data-points on the manifold must be distinctly reconstructed
- The contractive pressure is counteracted by the reconstruction gradient in the directions tangent to the manifold
- Can be interpreted roughly as curvature regularization

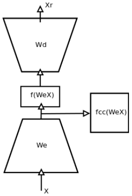
- The goal is to obtain a good representation (parameterization) of the data-manifold
- The reconstruction objective ensures that points on the manifold are well reconstructed (pushes down energy or raises probability)
- Without any other constraints, it often means that points that are off the data manifold are also well reconstructed
- This is not an issue when maximizing log-likelihood



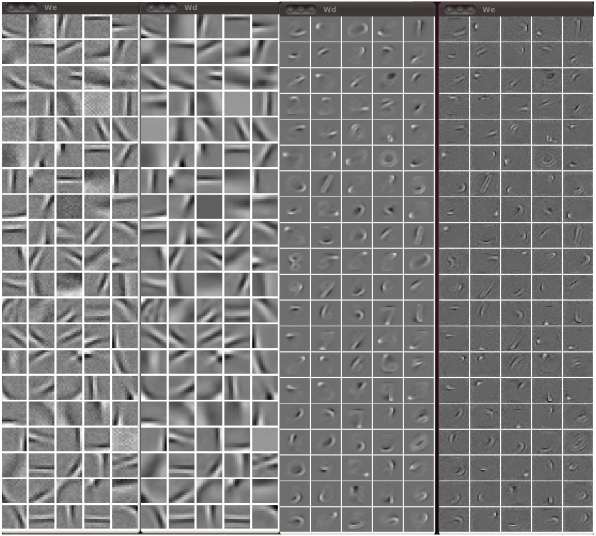
Saturating Auto-Encoder (SAE)

- Inspired by the CAE, we introduce a penalty on activations outside the saturated region of the nonlinearity

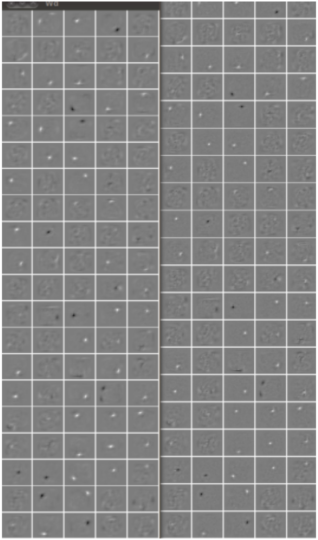




SAE: *shrink*() nonlinearity



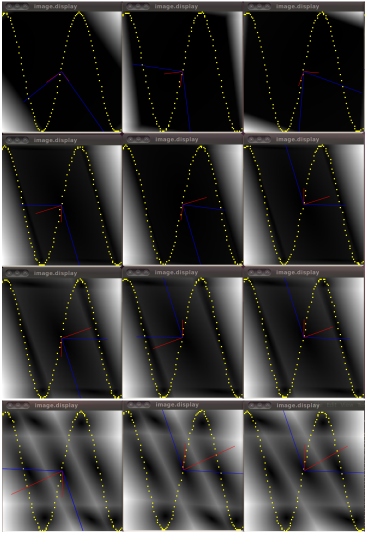
SAE: *sat_linear*() nonlinearity



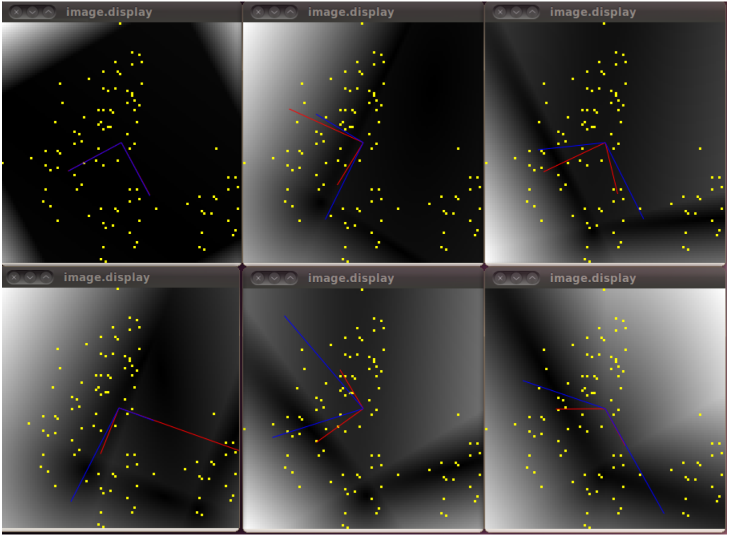
SAE: *sat_linear*() nonlinearity

SAE: *sat_linear*() nonlinearity

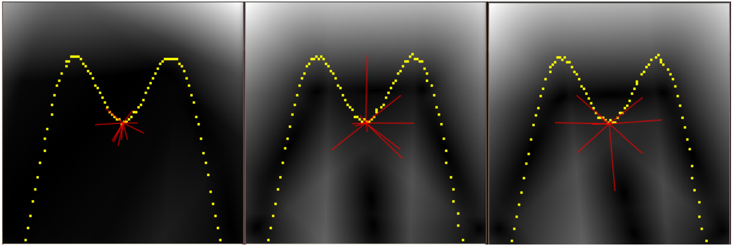
Reconstruction error surface for $\eta = 0, 0.05, 0.1, 0.2$



SAE: *sat_linear*() nonlinearity



SAE: *sat_linear*() nonlinearity



SAE: $shrink_+()$ nonlinearity