# Summary of (Baldi and Hornik, 1989)

We're interested in the energy landscape of a three layer linear neural network applied to least squares regression. The input dimensionality is $m$, the hidden dimensionality is $p$ and the output dimensionality is $n$, with weights $B$ taking us from input to the hidden layer, and weights $A$ taking us from the hidden layer to the output. The loss function we're interested in is then:

$$E(A,B) = \sum_t \|y_t - ABx_t\|_2^2 = \|Y - ABX\|_F^2.$$

Let $W = AB$. It's instructive to consider the case when $W$ is unconstrained by this relation. We then have

$$E(W) = \|Y - WX\|_F^2,$$

and setting the gradient with respect to $W$ to zero we get

$$W^* = YX^T(XX^T)^{-1} = \Sigma_{YX}\Sigma_{XX}^{-1}.$$

Our estimate of $Y$ is then

$$Y^* = W^*X = \Sigma_{YX}\Sigma_{XX}^{-1}X,$$

for which

$$\Sigma_{Y^*Y^*} = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

Because this covariance will come up repeatedly, we'll refer to it simply as $\Sigma$.

Returning to the case of interest where $W = AB$, we characterize the critical points of $E(A,B)$ by setting its derivatives with respect to $A$ and $B$ to zero. The derivatives are:

$$\nabla_A E = -2(Y - ABX)X^T B^T, \quad \nabla_B E = -2A^T(Y - ABX)X^T.$$

Setting these to zero, we have that at a critical point

$$\Sigma_{YX}B^T = AB\Sigma_{XX}B^T \tag{1}$$

$$A^T\Sigma_{YX} = A^TAB\Sigma_{XX} \tag{2}$$

Assuming $A^TA$ is invertible, equation 2 gives us $B$ in terms of $A$:

$$B(A) = (A^TA)^{-1}A^T\Sigma_{YX}\Sigma_{XX}^{-1},$$

and multiplying both sides on the left by $A$ we get

$$W = P_A\Sigma_{YX}\Sigma_{XX}^{-1},$$

where $P_A$ is the projection on the span of $A$. Multiplying on the right by $\Sigma_{XY}$ we get

$$W\Sigma_{XY} = P_A\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} = P_A\Sigma.$$

We eigendecompose $\Sigma$ into

$$\Sigma = U\Lambda U^T$$

and then write $P_A$ in terms of $P_{U^TA}$ as

$$P_{U^TA} = U^T P_A U \implies P_A = UP_{U^TA}U^T.$$

From $P_A\Sigma = \Sigma P_A$ we have

$$UP_{U^TA}U^T U\Lambda U^T = U\Lambda U^T UP_{U^TA}U^T \implies P_{U^TA}\Lambda = \Lambda P_{U^TA}.$$

Assuming the eigenvalues of $\Sigma$ are greater than zero and distinct, this relation implies that $P_{U^TA}$ is diagonal, because equating the elements of the left and right hand sides we get

$$LHS_{ij} = P_{U^TA,ij}\Lambda_j = RHS_{ij} = P_{U^TA,ij}\Lambda_i \implies P_{U^TA,ij} = 0 \text{ when } i \neq j.$$

Then because $P_{U^TA}$ is a projection operator, its diagonal elements must be either one or zero, and because it's a projection into a $p$-dimensional space, exactly $p$ of the diagonal elements must be one, but we are free to choose any of $\binom{n}{p}$ $p$-subsets of the $n$ elements. Call one such subset $I$. We then have

$$P_A = UP_{U^TA}U^T = U_I U_I^T,$$

so projection onto the span of $A$ is equivalent to projecting on to the span of $U_I$. This in turn implies that

$$A = U_I C, \quad B = C^{-1}U_I^T \Sigma_{YX}\Sigma_{XX}^{-1}, \quad W = U_I U_I^T \Sigma_{YX}\Sigma_{XX}^{-1} = U_I U_I^T W^*,$$

where $C$ is some invertible matrix. Hence the critical points are, up to a factor of $C$, indexed by taking size-$p$ subsets of the eigenvalues of $\Sigma$. It's then intuitive that the subset which takes the top $p$ eigenvalues is the global minimum, and the others are either local-minima or saddle points. To establish this, we first compute the energy of a critical point. We first note that by by multiplying both sides of equation 1 by $A^T$ we get

$$\Sigma_{YX}W^T = W\Sigma_{XX}W^T = W\Sigma_{YX}.$$

Then

$$\begin{aligned}
E(A,B) &= tr(\Sigma_{YY}) - 2tr(W\Sigma_{XY}) + tr(W\Sigma_{XX}W^T) \\
&= tr(\Sigma_{YY}) - tr(W\Sigma_{XY}) \\
&= tr(\Sigma_{YY}) - tr(U_I U_I^T \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}) \\
&= tr(\Sigma_{YY}) - tr(U_I U_I^T \Sigma) \\
&= tr(\Sigma_{YY}) - tr(U_I^T \Sigma U_I) \\
&= tr(\Sigma_{YY}) - tr(U_I^T U\Lambda U^T U_I) \\
&= tr(\Sigma_{YY}) - tr(\Lambda_I).
\end{aligned}$$

This shows that the subset $I$ which includes the top $p$ eigenvalues is the global minimum (when the eigenvalues are distinct, as we've assumed here). It remains to be shown whether the remaining $\binom{n}{p} - 1$ critical points are local minima or saddles. By perturbing their $A$ matrices in the direction of the missing eigenvectors we can show that local directions exist that reduce the energy, proving that the remaining critical points are all saddles. Thus we've characterized the landscape (up to a factor $C$).