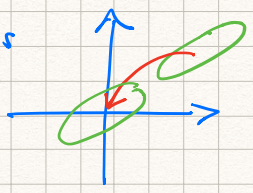
Batch normalization

The idea is to decorrelate the inputs to every layer in a NN.

Decorrelation requires a translation of all vectors

$$\vec{x} \to \vec{x} - \vec{\mu}$$



by subtracting the mean. It requires also a rotation and scaling:

$$\underbrace{\Sigma}_{\text{covariance matrix}} = U^T D U$$

U = rows are the eigenvectors of Σ

$$D = \begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & \cdots & \lambda_n \end{pmatrix}$$

λ_i: eigenvalues, represent the variance along the principal components

Decorrelation

$$(\Sigma^{-1})^{1/2} (\vec{x} - \vec{\mu})$$

Mahalanobis distance is measured for decorrelated vectors

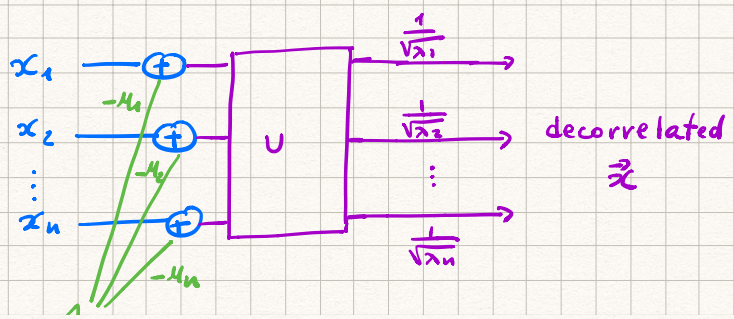
$$(x-\mu)^T (\Sigma^{-1})^{1/2} (\Sigma^{-1})^{1/2} (x - \vec{\mu})$$

$$= (x-\mu)^T \Sigma^{-1} (x - \vec{\mu})$$

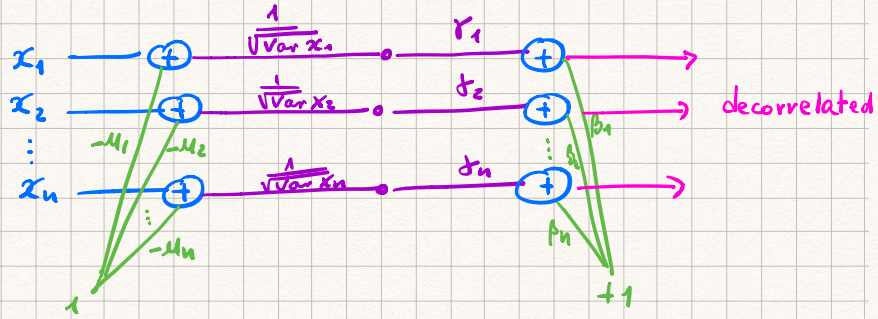
and $(\Sigma^{-1})^{1/2} = (U^T D^{-1} U)^{1/2} = U^T D^{1/2} U$

$$= U^T \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & & \\ & \frac{1}{\sqrt{\lambda_2}} & \\ & \cdots & \end{pmatrix} U$$

Batch normalization requires this network

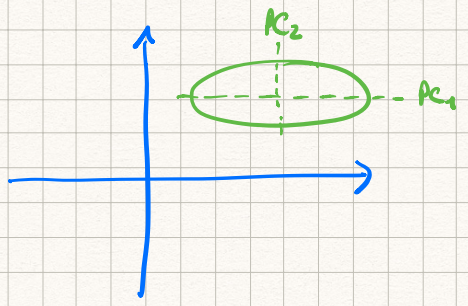


The authors of the BN paper simplify the computation to



The constants γ_i, β_i can "undo" the decorrelation step. They are learned in the BP step, during error minimization. The idea is that they can "mitigate" errors introduced by decorrelating.

The implicit assumption used by the simplification is that the principal components of the data are parallel to the axes:



The mean and variance of the input vectors is computed for each minibatch.

During inference the mean and variance of the input to each layer is the one for the complete population.