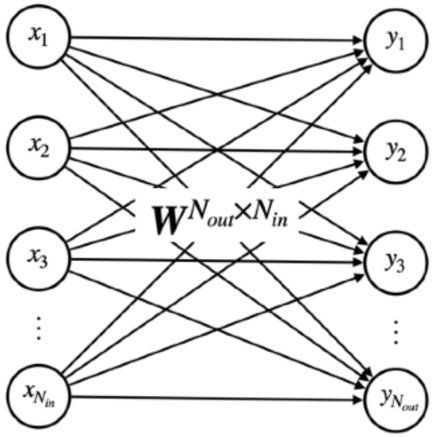
Method 2

Compressing the Linear Layer

The original layer

$$\boldsymbol{y}^{N_{out}} = \boldsymbol{W}^{N_{out} \times N_{in}} \boldsymbol{x}^{N_i n} + \boldsymbol{b}^{N_{out}}$$

The original layer



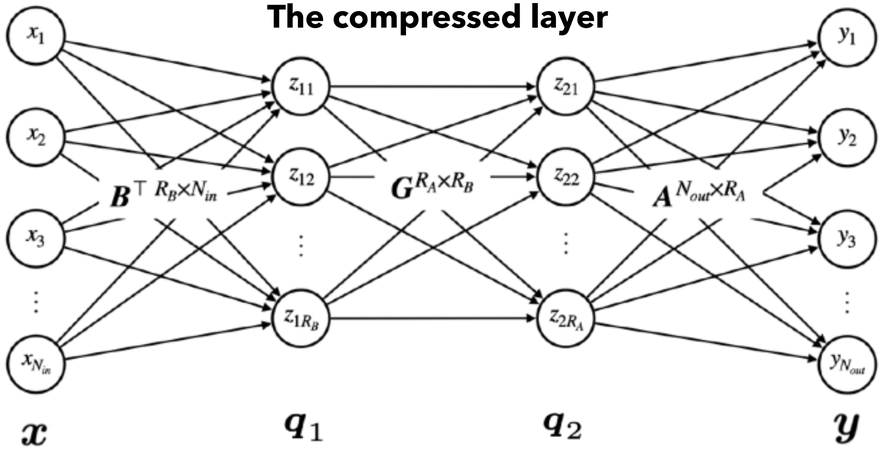
Tucker-2 approximation of weight matrix

$$\boldsymbol{W} \approx \boldsymbol{G} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} = \boldsymbol{A}\boldsymbol{G}\boldsymbol{B}^\top$$

The compressed layer

$$\boldsymbol{y}^{N_{\text{out}}} \approx \boldsymbol{A}^{N_{\text{out}} \times R_A} \underbrace{\boldsymbol{G}^{R_A \times N_{\text{in}}} \underbrace{\boldsymbol{B}^{\top R_B \times N_{\text{in}}} \boldsymbol{x}^{N_{\text{in}}}}_{\boldsymbol{q}_1}}_{\boldsymbol{q}_2} + \boldsymbol{b}^{N_{\text{out}}}$$

The compressed layer



Method 2 rank selection

Variational Bayesian Matrix Factorization

Acts as a good heuristic, but not guaranteed to yield the optimal ranks for the master problem

Given a matrix $\boldsymbol{V}$ it is assumed:

$$\boldsymbol{V}^{L\times M} = \boldsymbol{U}^{L\times M} + \boldsymbol{E}^{L\times M}$$

Goal is to find matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ such that

$$\boldsymbol{U} = \boldsymbol{B}\boldsymbol{A}^{\top}$$

Probabilistic model of $\boldsymbol{V}$ gives the posterior distributions of $\boldsymbol{A}$ and $\boldsymbol{B}$

Generally a non-convex problem, but analytical solution have been found