Training NNs via the Augmented Lagrangian Method (ALM)

NN model:

$$\Phi(W_0,\dots,W_N;z)=W_N\Phi(W_{N-1}\cdots\Phi(W_0x)\cdots)$$

NN training: Given data (z_0^j,y^j), $j=1,\dots,m$, solve

$$\min_{W_0,\dots,W_N}\frac{1}{2}\sum_{j=1}^{m}\|\Phi(W_0,\dots,W_N;z_0^j)-y^j\|_2^2$$

Introducing dynamics $z_{a+1}=\Phi(W_a z_a)$ we can express it as an optimal control problem

$$\min\frac{1}{2}\sum_{j=1}^{m}\|W_N z_N^j-y^j\|_2^2$$

$$\text{s.t. } z_{a+1}^j=\Phi(W_a z_a^j),\quad a=0,\dots,N-1,\ j=1,\dots,m$$

In abstract terms

$$\min\frac{1}{2}\|F(x)\|_2^2$$

$$\text{s.t. } h(x)=0$$

where $F(x)=(W_N z_N^1-y^1,\dots,W_N z_N^m-y^m)$

$h(x)=(h^1(x),\dots,h^m(x))$

$h^j(x)=(z_1^j-\Phi(W_0 z_0^j),\dots,z_N^j-\Phi(W_{N-1}z_{N-1}^j))$

$x=(W_0,\dots,W_N,(z_1^j,\dots,z_N^j)_{j=1}^m)$

$$\min \frac{1}{2}\|F(x)\|_2^2$$
$$\text{s.t.} \quad h(x)=0$$

Augmented Lagrangian

$$\mathcal{L}_\beta(x,y) = \frac{1}{2}\|F(x)\|_2^2 + \langle y, h(x)\rangle + \frac{\beta}{2}\|h(x)\|_2^2$$
$$= \frac{1}{2}\|F(x)\|_2^2 + \frac{\beta}{2}\|h(x)+y/\beta\|_2^2 - \frac{1}{2\beta}\|y\|_2^2$$
$$= \frac{\beta}{2}\left\|\begin{bmatrix} F(x)/\sqrt{\beta} \\ h(x)+y/\beta \end{bmatrix}\right\|^2$$

ALM: Given y^0, iterate

1. Find x^k s.t. $\|\nabla_x \mathcal{L}_{\beta_k}(x^k,y^k)\|_2 \le \epsilon_k$
2. Update multipliers $y^{k+1} = y^k + \beta_k h(x^k)$

Step 1 amounts to solving a nonlinear least-squares problem:
$$\min \frac{1}{2}\|F_{\beta_k}(x,y^k)\|_2^2$$
where $F_\beta(x,y) = \begin{bmatrix} F(x)/\sqrt{\beta} \\ h(x)+y/\beta \end{bmatrix}$

Can use Levenberg-Marquardt (LM)

Questions:

1. How can we solve the linear least-squares problem efficiently in the LM method?
   It has a lot of structure: exploit it!
2. How can we adjust penalty parameters β_k and tolerances ϵ_k? Check the literature!
3. Other methods for solving step 1?