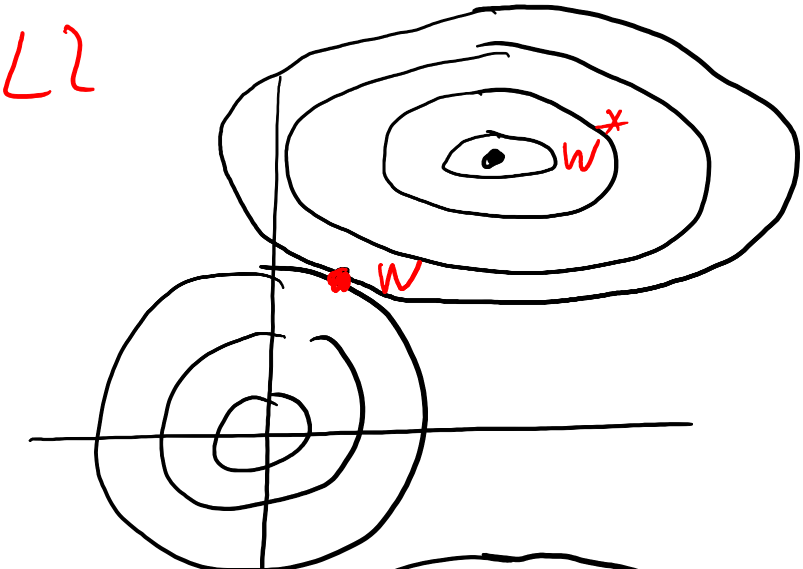
L7 EC500K 2/23

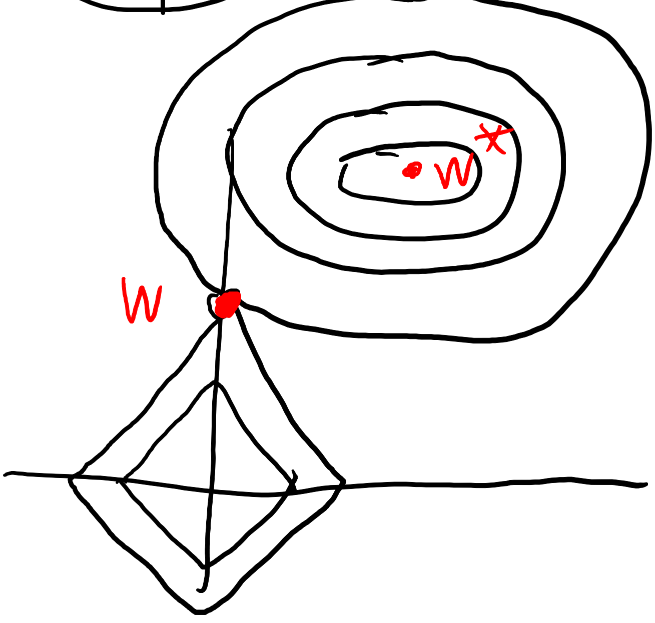
- DL works well with high dimension
- Deeper network requires narrower layer
- Regularize bias will underfit model
- $W_{n+1} = \overset{L_2}{W_n - \varepsilon(\alpha W_n + \nabla_{W_n} J)} = \overset{L_1}{W_n - \varepsilon(\alpha\, \mathrm{sign}(W_n) + \nabla_{W_n} J)}$

  $= (1-\epsilon\alpha)W - \epsilon \nabla_{W_n} J = (1-\epsilon\alpha)\mathrm{sign}(W_n) - \epsilon \nabla_{W_n} J$
- L2



w* overfits

L1



- Minibatch SGD:

  10000 samples, 200 batches.

  SGD compute 200 more calculations,

  but reduce error by factor of $\sqrt{50}$

- Using GPU, batch size should be in size of 2^k, ie 32, 64, 128, 256
- Batch is also considered as regularizer
- Hessian matrix requires more samples per batch, but size of W can be too large!!!

  So we don't use 2nd order method.

- Read on momentum & initialization
- Initializing weights should break symetry.

  $W_0 \sim N(0, \sigma^2)$ & $U(-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}})$ & $U(-\frac{1}{\sqrt{m+n}}, \frac{1}{\sqrt{m+n}})$

- Initializing bais to match marginal distribution of the data; bias should not be zero.