Regularization

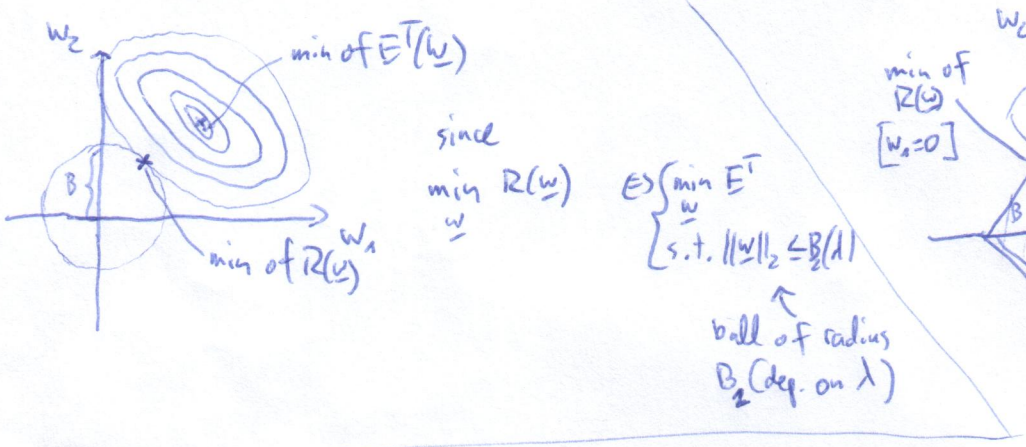
a) Approach to overfitting reduction: L_1/L_2 regularization.

if optimization of E^T yields too large $\hat{E}^G \Rightarrow$ regularize by optimizing risk $R(\underline{w}) = E^T(\underline{w}) + \lambda E^R(\underline{w})$ instead!

E^T: training error; λ: regularization strength (hyperparameter); E^R: regularization term, $\lambda > 0$

L_2 norm regularization: "weight decay": $E^R(\underline{w}) = \frac{1}{2p}\|\underline{w}\|_2^2 = \frac{1}{2p}\sum_{i=1}^{d} w_i^2$

- → standard regularizer
- → differentiable (gradient descent straight): $\frac{\partial E}{\partial \underline{w}} = \frac{\partial E^T}{\partial \underline{w}} + \lambda \frac{\partial E^R}{\partial \underline{w}}$, $\frac{\partial E^R}{\partial \underline{w}} = \frac{1}{p}\underline{w}$
- → robust to noise (distributed model: redundancies in weights allowed)
- → non-sparse



L_1 norm regularization: "Lasso" / "sparsify": $E^R(\underline{w}) = \frac{1}{p}\|\underline{w}\|_1 = \frac{1}{p}\sum_{i=1}^{d}|w_i|$

- → sparse (good for interpretable models, e.g. in medical applications)
- → non-differentiable at $w_i = 0$ $\forall i$ $\Rightarrow$ gradient descent not straight applicable (but solved via subgradient / least angle regression)
- → more sensitive to ~~outliers~~ noise since much less redundancies

b) Analytical solution for linear neuron of quadratic cost function and L_2 regularization.

Cost function $R(\underline{w}) = \frac{1}{p}\sum_{\alpha=1}^{p} \underbrace{\frac{1}{2}\left(\underline{w}^T \underline{x}^{(\alpha)} - y_T^{(\alpha)}\right)^2}_{E^{(\alpha)}} + \lambda \frac{1}{2p}\underline{w}^T\underline{w}$

$E^{(\alpha)}$: quadratic cost of data pt. α

Gradient $\underline{g} = \frac{1}{p}\left(\underline{X}\underline{X}^T\underline{w} - \underline{X}\underline{y}^T\right) + \frac{\lambda}{p}\underline{w}$ $\overset{\underline{g}=\underline{0}}{\Rightarrow}$ $\underline{w}^* = \left(\underline{X}\underline{X}^T + \lambda \underline{I}\right)^{-1}\underline{X}\underline{y}^T$ ($\underline{I}$: identity matrix)

Hessian $\underline{H} = \frac{1}{p}\left(\underline{X}\underline{X}^T + \lambda\underline{I}\right)$ still positive definite $\Rightarrow$ $\underline{w}^*$ unique global minimum

$\left[\underline{X} = (\underline{x}^{(1)}, \dots, \underline{x}^{(p)})$ data columnwise$\right]$