Regularization

$\min_w \mathcal{L}(w)$ $+\lambda \|w\|_F^2$

⇒ weight decay

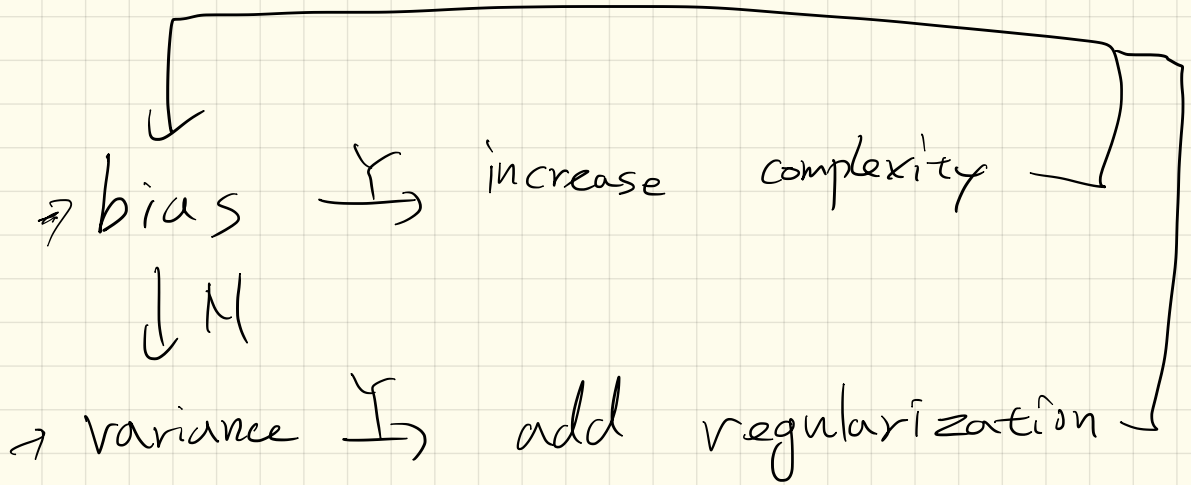
$$w(t+1) = (1-\alpha) \cdot w(t) - \gamma \nabla \mathcal{L}(w(t))$$

⇒ dropout

⇒ Data Augmentation. (⇒ next time)

⇒ Batch Norm.

⇒ SGD ⇒ Early stopping.


bias
Y
increase complexity
N
variance
Y
add regularization

$$\hat{m}^{[0]} = 0$$

$$\hat{m}^{[1]} = (1-\eta) \cdot \hat{m}^{\{1\}[0]}$$

$$\hat{m}^{[2]} = (1-\eta) \cdot \hat{m}^{\{2\}[1]} + \eta \cdot (1-\eta)\, \hat{m}^{\{1\}[1]}$$

$$\hat{m}^{[3]} = (1-\eta)\, \hat{m}^{\{3\}[2]} + (1-\eta)\, \eta\, \hat{m}^{\{2\}[2]} + (1-\eta)\, \eta^2\, \hat{m}^{\{1\}[2]}$$

$$\vdots$$

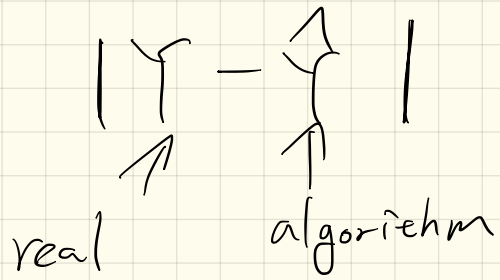
$$\hat{m}^{[l]} = (1-\eta)\, \hat{m}^{\{W\}[l]} + (1-\eta) \cdot \eta\, \hat{m}^{\{1-W\}[l]} + \cdots + (1-\eta) \cdot \eta^{1-W}\, \hat{m}^{\{1\}[l]}$$

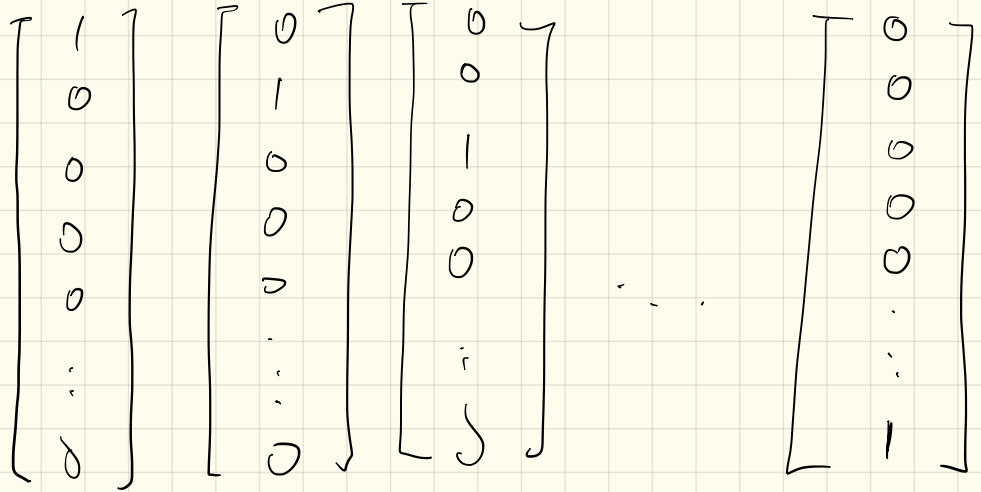
$$= (1-\eta^W)\, \frac{\hat{m}^{\{w\}[l]} + \eta\, \hat{m}^{\{1-W\}[l]} + \cdots + \eta^{W-1}\, \hat{m}^{\{1\}[l]}}{1 + \eta + \cdots + \eta^{1-W}}$$

MNIST

0
1
2
3
4
5
6
7
8
9

10 classes





class 0 class 1 . . . class 9

$$\hat{Y} = \begin{bmatrix} P(Y=0|X) \\ P(Y=1|X) \\ \vdots \\ P(Y=9|X) \end{bmatrix}$$

Cross entropy

Cross entropy ?

$$\hat{Y} = \begin{bmatrix} P(Y=0|X) \\ \vdots \\ P(Y=9|X) \end{bmatrix} \Rightarrow \log \hat{Y} = \begin{bmatrix} \log P(Y=0|X) \\ \vdots \\ \log P(Y=9|X) \end{bmatrix}$$

output of our algorithm.

$$Y^T \cdot (\log \hat{Y}) = -\log P_{W.b.}(Y = Y^* | X)$$

Y: one hot vector of true label