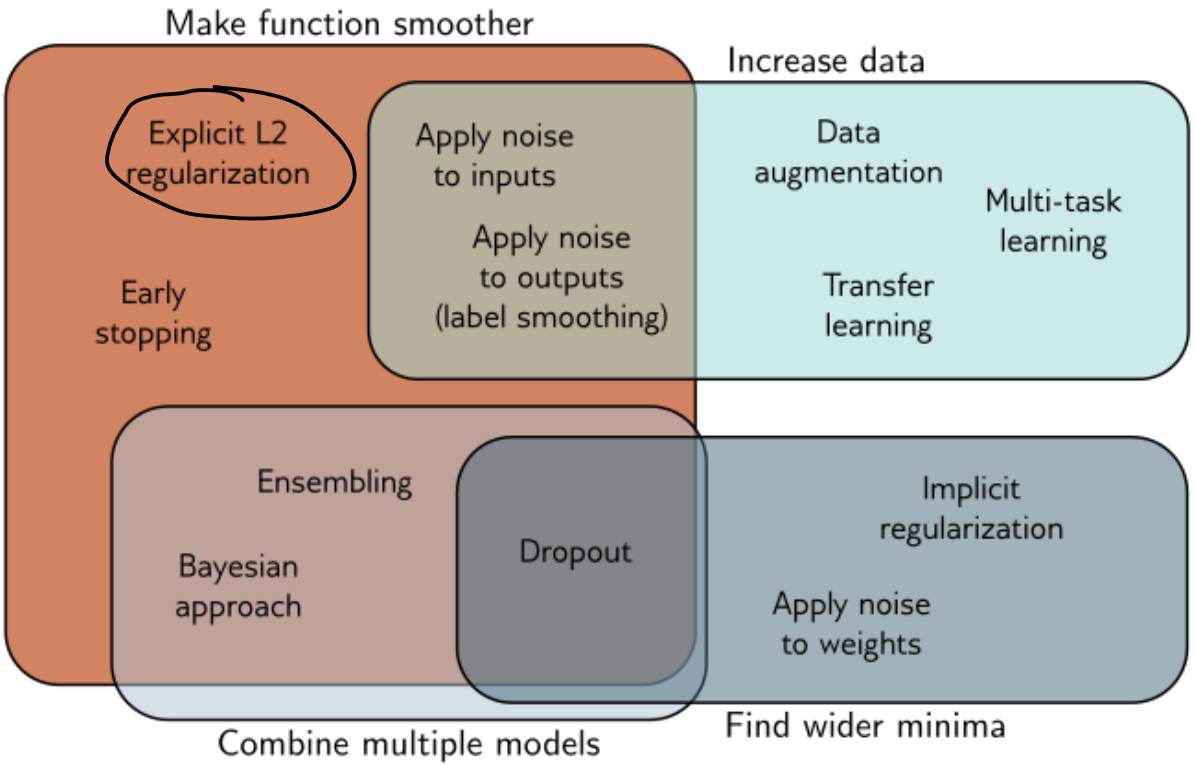
Regularization (Chapter 9 of UDL book)




Regularization

$$\underset{\theta}{\operatorname{argmin}}\; L(D;\theta) + \underbrace{\lambda R(\theta)}_{\text{Regularizer}}$$

$$\exp(-\lambda R(\theta))$$

$$\equiv \underset{\theta}{\operatorname{argmax}}\; P(\theta|D) = P(D|\theta)\,P(\theta)$$

L-2 regularization

$$\underset{w}{\operatorname{argmin}}\; \underbrace{\|y - Xw\|_2^2}_{\text{Loss function}} + \lambda \underbrace{\|w\|_2^2}_{\text{L-2 norm of parameters}}$$

$$\underbrace{L(D,\theta)}$$

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \| y - \underbrace{f(x;\theta)}_{\hat{y}} \| \rightarrow \lambda \| \theta \|_2^2$$

$$\hookrightarrow \theta_{t+1} = \theta_t - \alpha_t \left( \frac{\partial L}{\partial \theta} + 2(\lambda)\theta_t \right)$$

Neural network call L-2 regularization as **weight decay**

$\sigma(Wx+b)$ ← not L-2 regularized

weight decay parameter $= \lambda$

$$\|\theta\|^2 = \theta^T \theta$$

$$= \theta^T \underset{n \times n}{I} \theta$$

$$\frac{\partial}{\partial \theta} \theta^T I_{n \times n} \theta = 2\theta$$

→ Parameters ↷ $\theta$ trained by gradient descent or SGD

the Loss function should be fixed

→ Hyper parameters

$\alpha_t$ , $\lambda$, Batch size

③ Contrastive divergence

① Grid search

$\lambda = [0.0, 0.01, 0.001, 0.1, 0.2, \ldots 0.9]$

a) Train ).

b) find validation accuracy

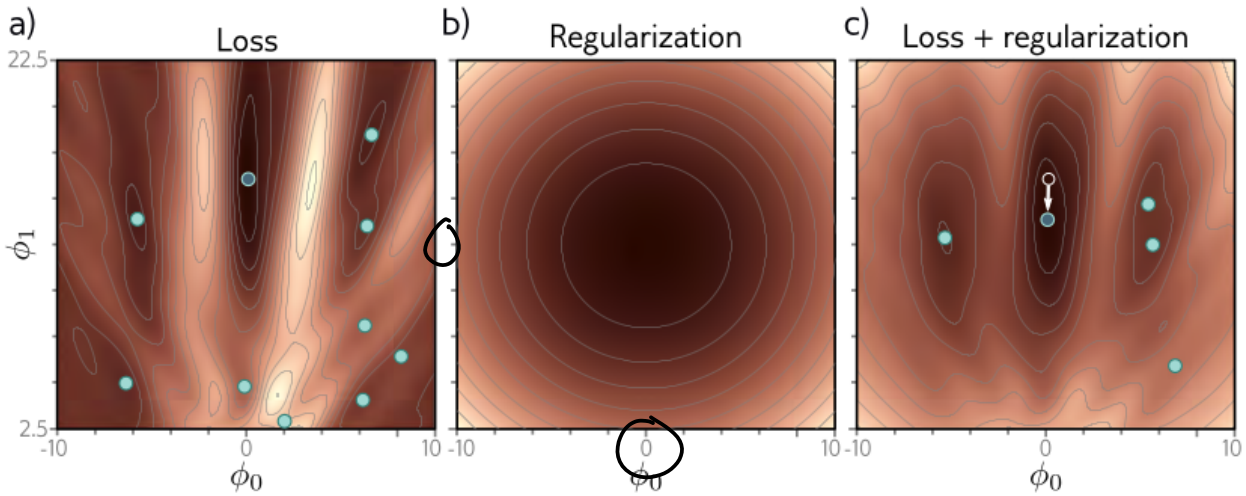
c) Pick the best set of hyper parameters

② Random search

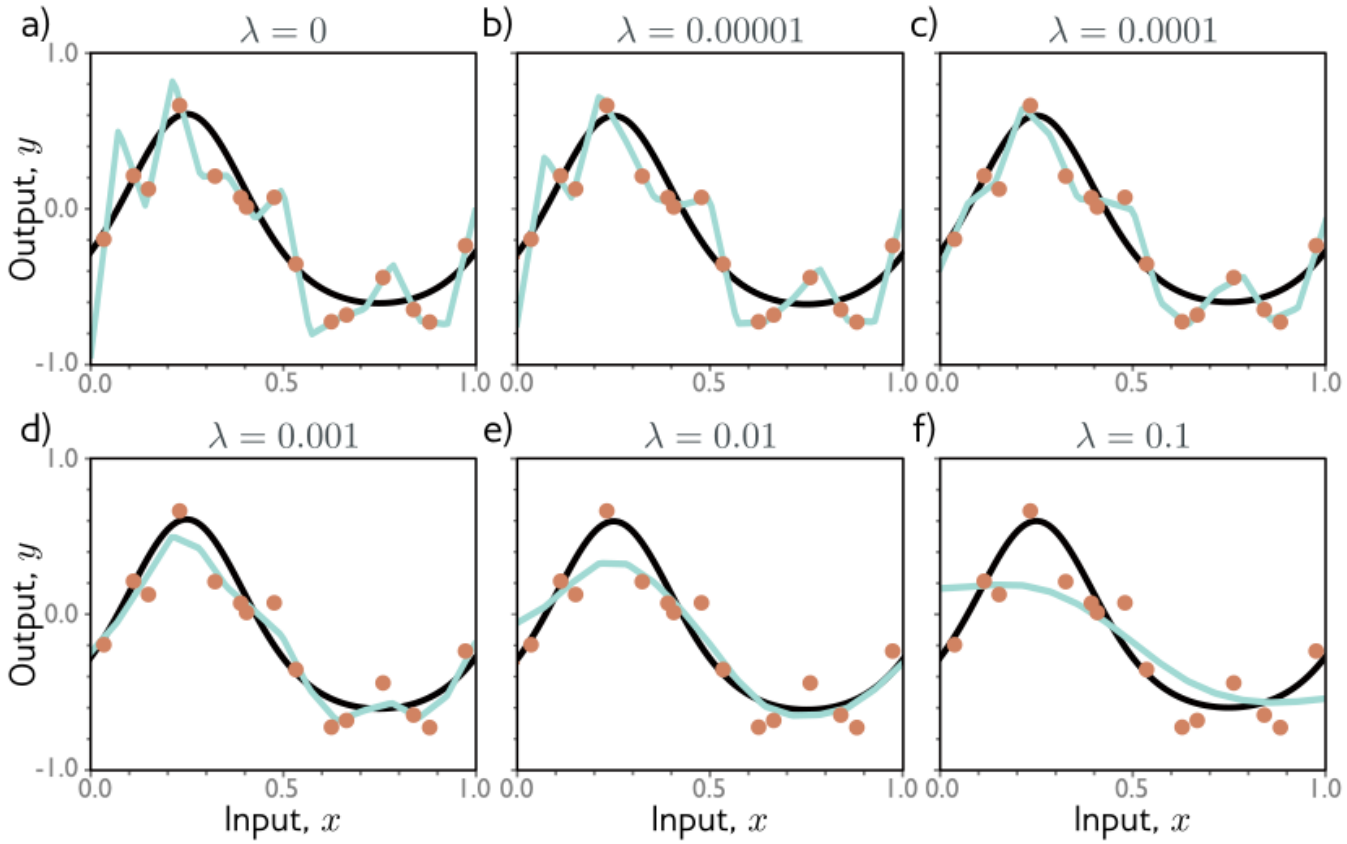
$\lambda \sim LL[0.0001, 0.99]$

a) Train (b) find validation (c) Pick the best

## Explicit regularization



## L2 Regularization




Regularization

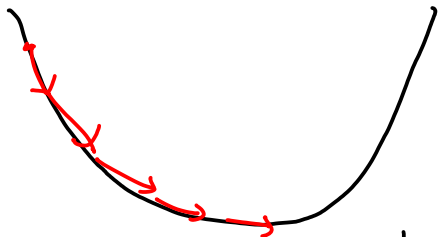
→ Explicit — L2 regularization
L-1 //
L-p //

→ Implicit regularization
→ Stochastic gradient descent → Batch size
→ Step size
→ Learning rate
→ Gradient descent → Step size
→ Learning rate
→ Heuristic regularization
→ Dropout etc

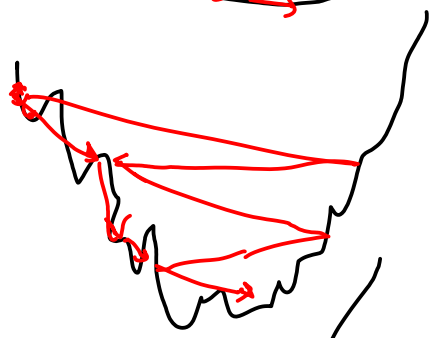
GD

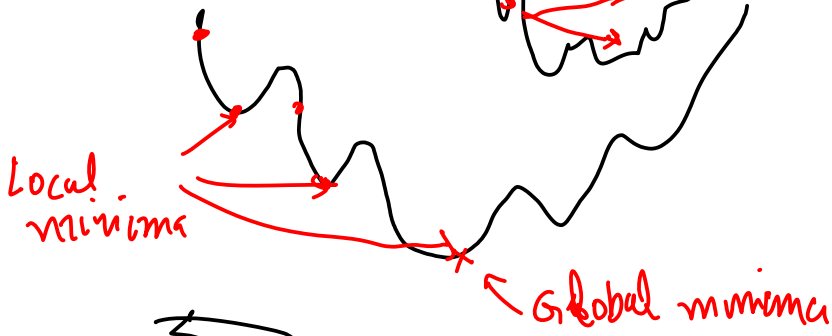
# Gradient descent induce regularization

$$\theta_{t+1} = \theta_t - \alpha \frac{dL}{d\theta}\Big|_{\theta=\theta_t}$$



Ideally GD should converge to Local Minima






## Gradient flow

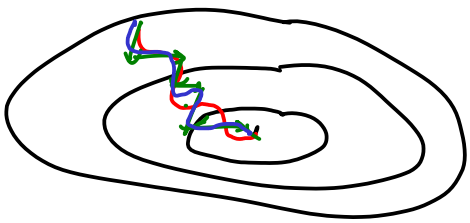
$$\frac{d\theta}{dt} = -\alpha \frac{dL}{d\theta}$$

$$\dot{\theta} = -g(\theta(t)) \quad g(\theta(t))$$

Ideal

Error

Actually: $\theta_{t+1} = \theta_t - g(\theta(t))$



Corrected Gradient flow

$$\dot{\theta} = -\tilde{g}(\theta(t)) = -g(\theta(t)) - \varepsilon g_1(\theta(t))$$

error term

$$-\frac{d\tilde{L}_{GD}}{d\theta} = \frac{dL}{d\theta} + (\varepsilon \text{term})$$

regularizer

$$\theta_{t+1} = \theta_t - \alpha \frac{\partial L}{\partial \theta}$$

By Taylor series expansion

$$\theta(t+\varepsilon) = \theta(t) + \varepsilon \frac{\partial \theta(t)}{\partial t} + \frac{\varepsilon^2}{2} \frac{\partial^2 \theta(t)}{\partial t^2} + \dots$$

$$\frac{\partial \theta}{\partial t}(t) = -\tilde{g}(\theta(t)) = -g(\theta(t)) - \varepsilon g_1(\theta(t)) \quad \text{①}$$

$$\frac{\partial^2 \theta}{\partial t^2}(t) = -\frac{\partial}{\partial \theta} g(\theta(t)) \frac{\partial \theta}{\partial t} - \varepsilon \left[ \frac{\partial}{\partial t} g_1(\theta(t)) \right] \quad \text{②}$$

$$\theta(t+\varepsilon) = \theta(t) - \varepsilon g(\theta(t)) - \varepsilon^2 g_1(\theta(t)) - \frac{\varepsilon^2}{2} \left[ \frac{\partial g(\theta(t))}{\partial \theta} \frac{\partial \theta}{\partial t} \right] - \left[ \frac{\varepsilon^3}{2} \frac{\partial}{\partial t} g_1(\theta(t)) \right]$$

① ② $O(\varepsilon^3)$

If $\varepsilon$ is very small we can ignore terms with $\varepsilon^3$

To make the equation as close as possible to

$$\theta(t+\varepsilon) = \theta(t) - \varepsilon g(\theta(t))$$

we need $$g_1(\theta(t)) = -\frac{1}{2} \frac{\partial g(\theta(t))}{\partial \theta} \frac{\partial \theta}{\partial t}$$

Corrected gradient flow

$$\dot{\theta} = -g(\theta) - \varepsilon g_1(\theta)$$

$$= -g(\theta) + \frac{\varepsilon}{2} \frac{\partial}{\partial \theta} g(\theta) \frac{\partial \theta}{\partial t}$$

$$= -\frac{\partial L}{\partial \theta} + \frac{\varepsilon}{2} \frac{\partial^2 L(\theta)}{\partial \theta^2} \frac{\partial \theta}{\partial t}$$

$$\frac{\partial \theta}{\partial t} = -g(\theta) = -\frac{\partial L}{\partial \theta}$$

$$= -\frac{\partial L}{\partial \theta} - \frac{\varepsilon}{2} \left[\frac{\partial^2 L(\theta)}{\partial \theta^2}\right] \frac{\partial L}{\partial \theta}$$

$$\dot{\theta} = -\frac{\partial}{\partial \theta} L_{GD} = -\frac{\partial}{\partial \theta}\left(L + \frac{\varepsilon}{4} \left\| \frac{\partial L}{\partial \theta} \right\|_2^2\right)$$

(scalar)

$$L_{GD} = L + \frac{\alpha}{4} \left\| \frac{\partial L}{\partial \theta} \right\|_2^2$$

Regularizer

Gradient descent implicitly introduces a regularizer of the size proportional to step size $\alpha$ and squared norm of gradients $\left\| \frac{\partial L}{\partial \theta} \right\|_2^2$

$$\frac{\partial L}{\partial \theta} = g$$

$$\frac{\partial}{\partial \theta} \|g\|^2 = 2 g \frac{\partial g}{\partial \theta}$$

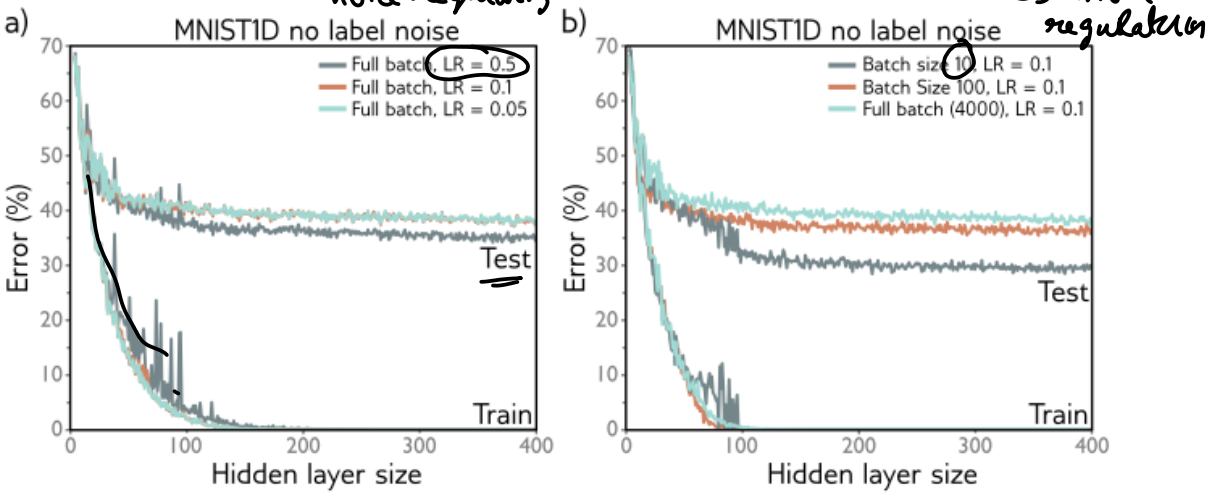
$$= 2 \frac{\partial L}{\partial \theta} \frac{\partial^2 L}{\partial \theta^2}$$

SGD introduces additional regularizer to GD that is proportional to the variance of batch gradients.

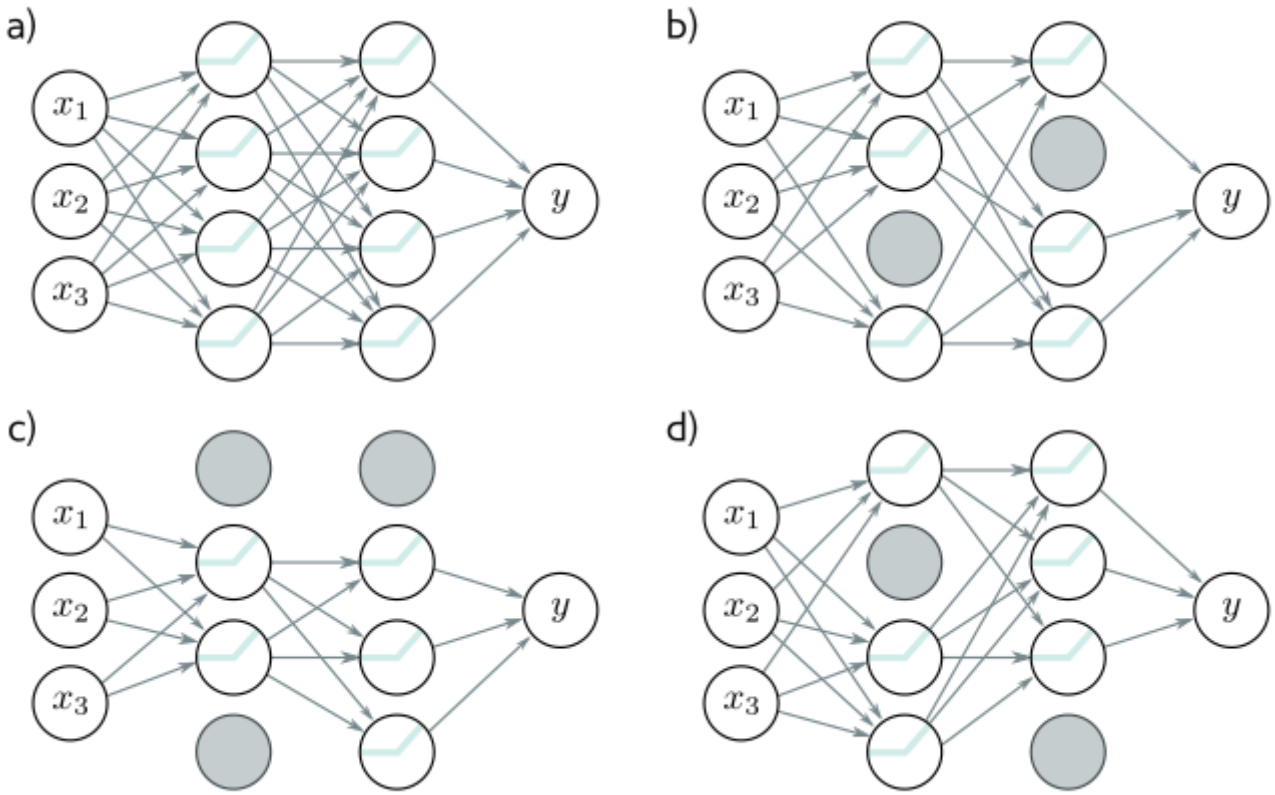
$$L_{SGD} = L_{GD} + \frac{\epsilon}{4B} \sum_{b=1}^{M} \left\| \frac{\partial L_b}{\partial \theta} - \frac{\partial L}{\partial \theta} \right\|_2^2$$

$$\begin{aligned} \tilde{L}_{SGD}[\phi] &= \tilde{L}_{GD}[\phi] + \frac{\alpha}{4B} \sum_{b=1}^{B} \left\| \frac{\partial L_b}{\partial \phi} - \frac{\partial L}{\partial \phi} \right\|^2 \\ &= L[\phi] + \frac{\alpha}{4} \left\| \frac{\partial L}{\partial \phi} \right\|^2 + \frac{\alpha}{4B} \sum_{b=1}^{B} \left\| \frac{\partial L_b}{\partial \phi} - \frac{\partial L}{\partial \phi} \right\|^2 . \end{aligned}$$

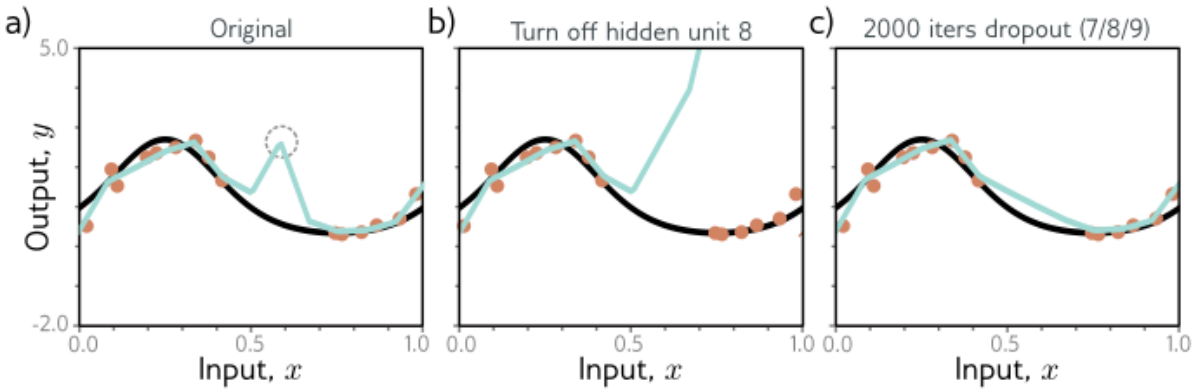
## Effect of batch size and learning rate



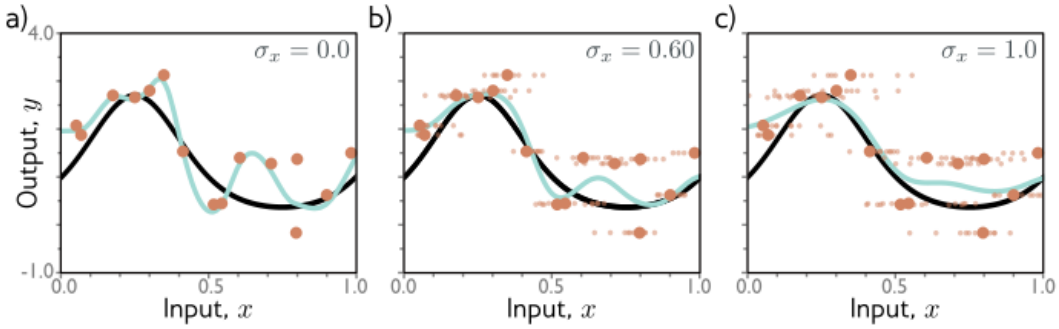
Dropout



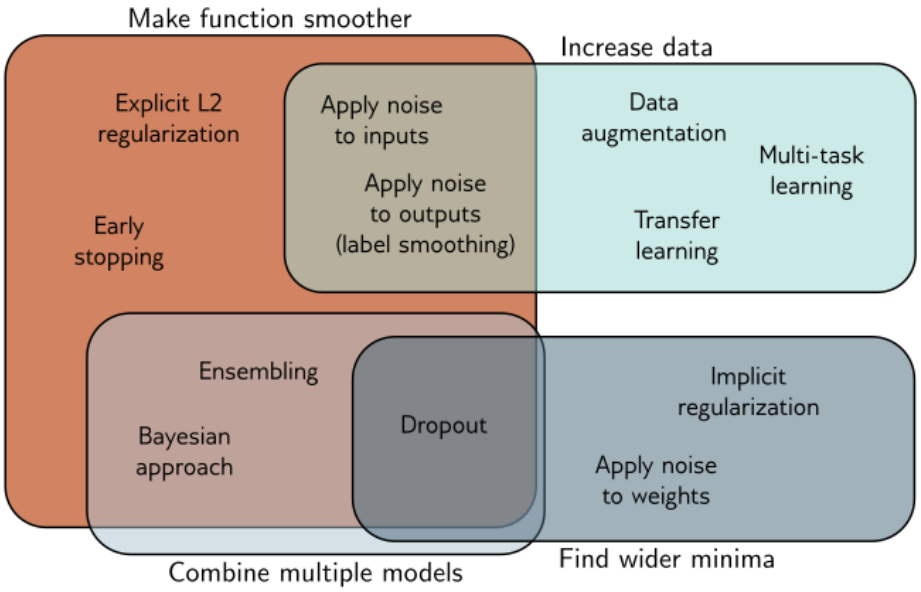
## Effect of dropout



## Adding noise to each batch



## Data augmentation


Make function smoother
Increase data
Explicit L2 regularization
Apply noise to inputs
Data augmentation
Multi-task learning
Apply noise to outputs (label smoothing)
Early stopping
Transfer learning
Ensembling
Implicit regularization
Dropout
Bayesian approach
Apply noise to weights
Combine multiple models
Find wider minima