CS7015 (Deep Learning) : Lecture 9

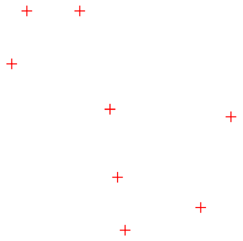
Regularization: Bias Variance Tradeoff, L2 regularization, Early stopping, Dataset augmentation, Parameter sharing and tying, Injecting noise at input, Ensemble methods, Dropout

February 10, 2017

Mitesh M. Khapra

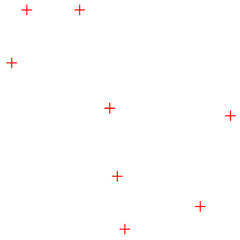
Department of Computer Science and Engineering
Indian Institute of Technology Madras

We will begin with a quick overview of bias, variance and the trade-off between them.



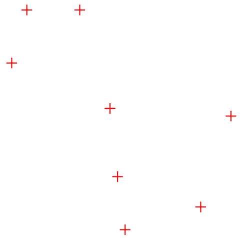
The points were drawn from a sinusoidal function (the true $f(x)$)

- Let us consider the problem of fitting a curve through a given set of points



The points were drawn from a sinusoidal function (the true $f(x)$)

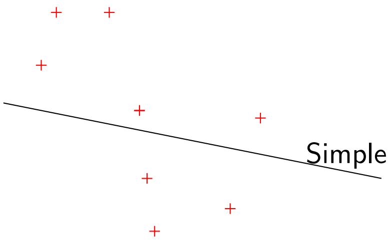
- Let us consider the problem of fitting a curve through a given set of points
- We consider two models :



The points were drawn from a sinusoidal function (the true $f(x)$)

- Let us consider the problem of fitting a curve through a given set of points
- We consider two models :

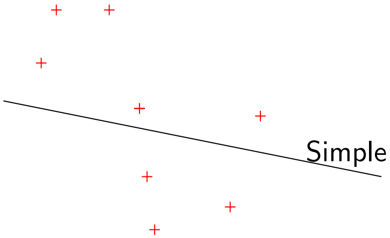
$$\underset{(degree:1)}{Simple} \quad y = \hat{f}(x) = w_1 x + w_0$$



The points were drawn from a sinusoidal function (the true $f(x)$)

- Let us consider the problem of fitting a curve through a given set of points
- We consider two models :

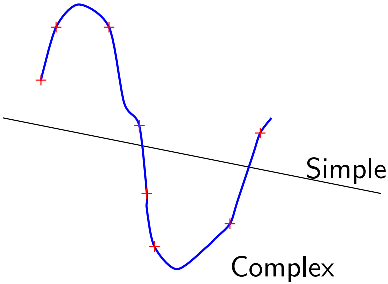
$$\underset{(\textit{degree}:1)}{\textit{Simple}} \quad y = \hat{f}(x) = w_1 x + w_0$$



The points were drawn from a sinusoidal function (the true $f(x)$)

- Let us consider the problem of fitting a curve through a given set of points
- We consider two models :

$$\begin{array}{cl} \textit{Simple} \\ (\textit{degree}:1) & y = \hat{f}(x) = w_1 x + w_0 \\ \textit{Complex} \\ (\textit{degree}:25) & y = \hat{f}(x) = \sum_{i=1}^{25} w_i x^i + w_0 \end{array}$$

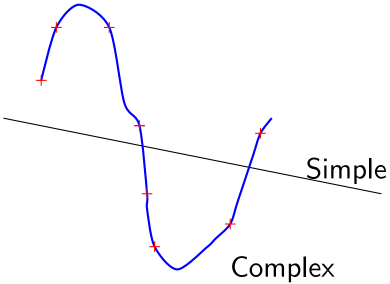


The points were drawn from a sinusoidal function (the true $f(x)$)

- Let us consider the problem of fitting a curve through a given set of points
- We consider two models :

$$\underset{(degree:1)}{\textit{Simple}} \quad y = \hat{f}(x) = w_1 x + w_0$$

$$\underset{(degree:25)}{\textit{Complex}} \quad y = \hat{f}(x) = \sum_{i=1}^{25} w_i x^i + w_0$$



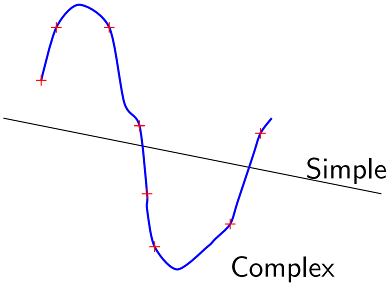
The points were drawn from a sinusoidal function (the true $f(x)$)

- Let us consider the problem of fitting a curve through a given set of points
- We consider two models :

$$\underset{(degree:1)}{\textit{Simple}} \quad y = \hat{f}(x) = w_1 x + w_0$$

$$\underset{(degree:25)}{\textit{Complex}} \quad y = \hat{f}(x) = \sum_{i=1}^{25} w_i x^i + w_0$$

- Note that in both cases we are making an assumption about how y is related to x. We have no idea about the true relation $f(x)$



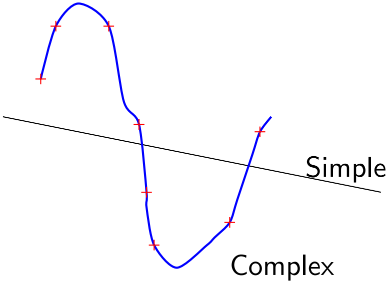
The points were drawn from a sinusoidal function (the true $f(x)$)

- Let us consider the problem of fitting a curve through a given set of points
- We consider two models :

$$\underset{(degree:1)}{Simple} \quad y = \hat{f}(x) = w_1 x + w_0$$

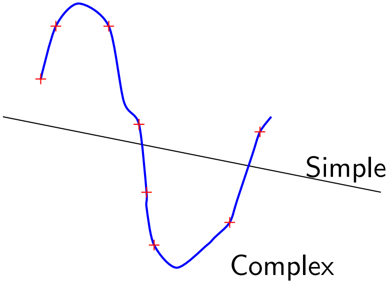
$$\underset{(degree:25)}{Complex} \quad y = \hat{f}(x) = \sum_{i=1}^{25} w_i x^i + w_0$$

- Note that in both cases we are making an assumption about how y is related to x. We have no idea about the true relation $f(x)$
- The training data consists of 100 points



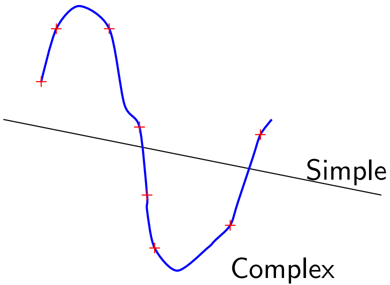
The points were drawn from a sinusoidal function (the true $f(x)$)

- We sample 25 points from the training data and train a simple and a complex model



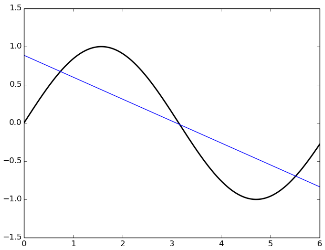
The points were drawn from a sinusoidal function (the true $f(x)$)

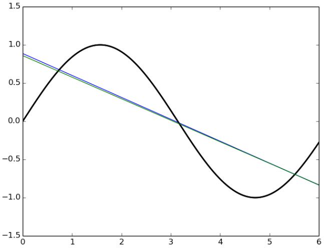
- We sample 25 points from the training data and train a simple and a complex model
- We repeat the process 'k' times to train multiple models (each model sees a different sample of the training data)

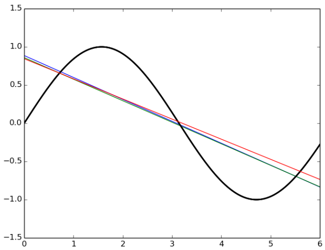


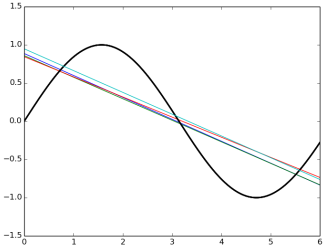
The points were drawn from a sinusoidal function (the true $f(x)$)

- We sample 25 points from the training data and train a simple and a complex model
- We repeat the process 'k' times to train multiple models (each model sees a different sample of the training data)
- We make a few observations from these plots

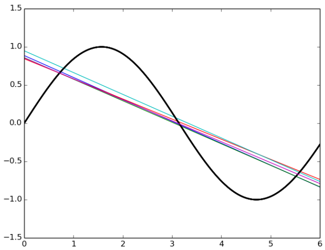

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

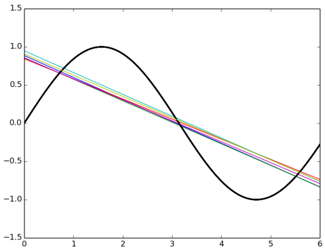

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6



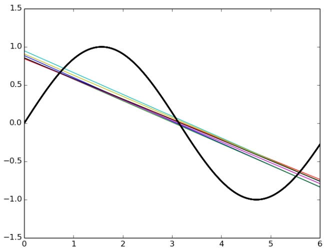
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

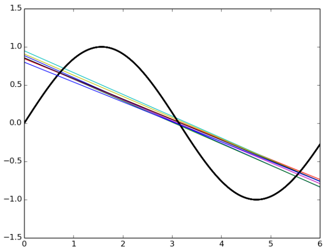


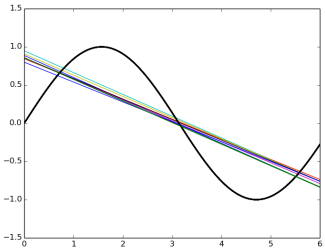
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

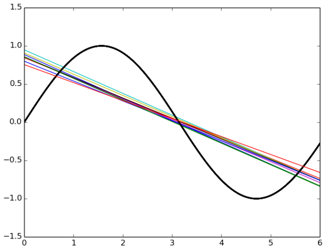


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

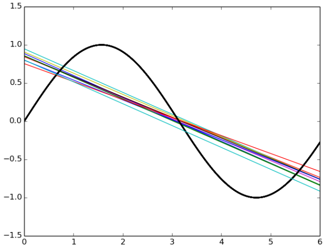

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

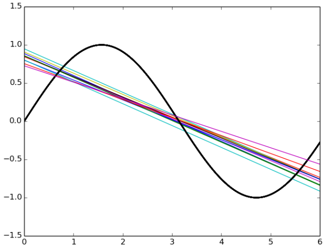

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

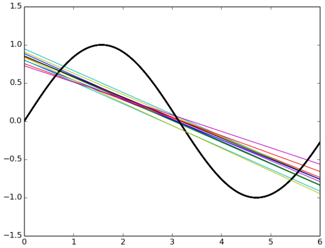

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6



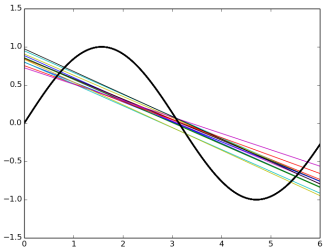
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

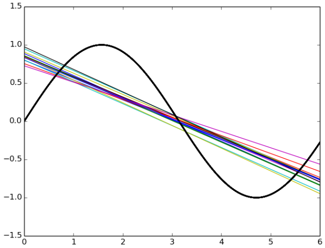

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

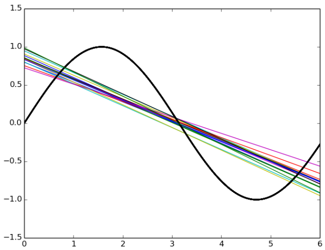

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

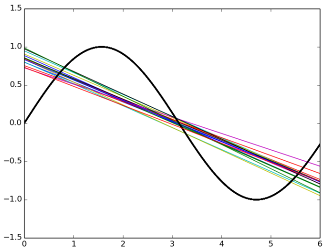


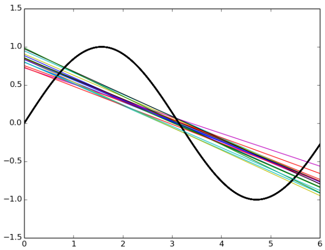
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

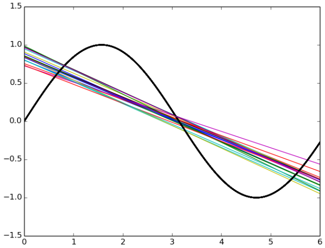

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

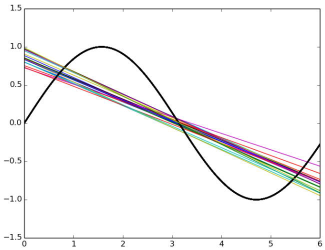

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

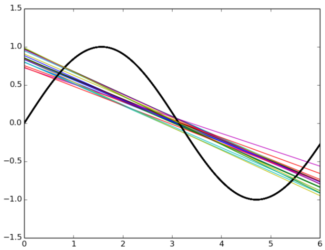

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6

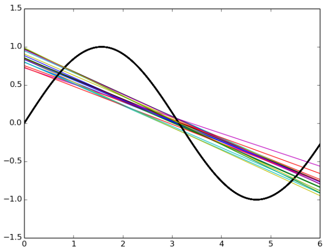

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6



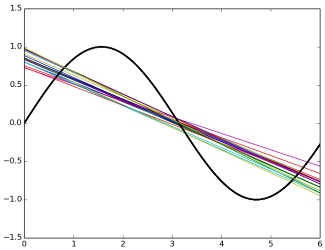
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5
6



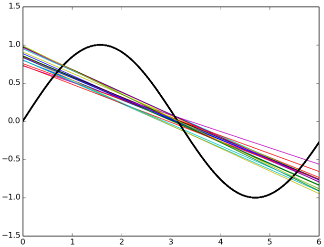
- Simple models trained on different samples of the data do not differ much from each other

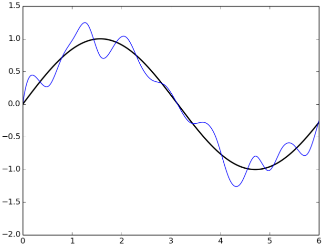


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)

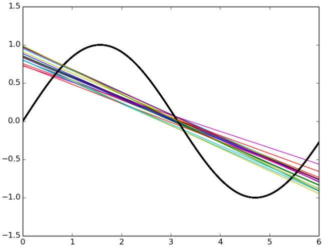


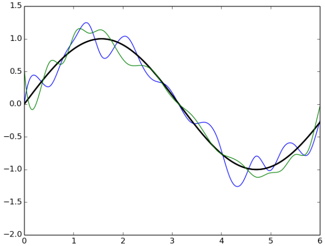
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



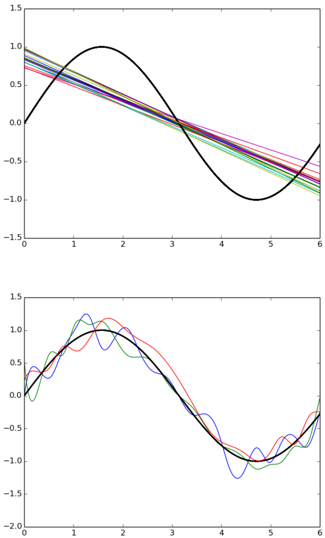


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)

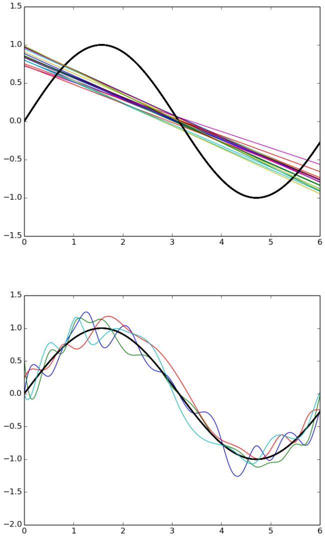




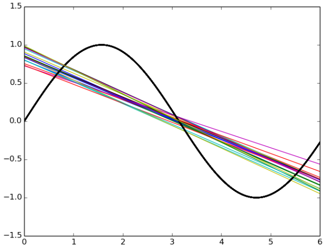
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)

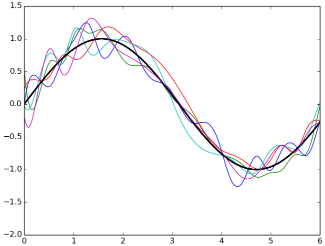


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)

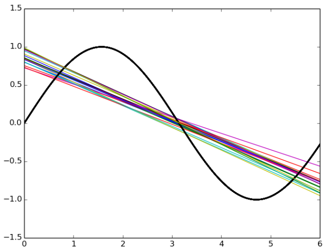


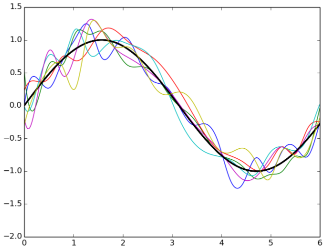
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



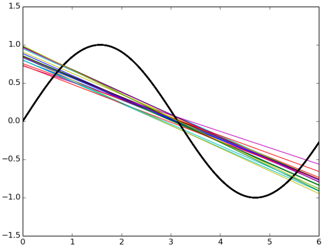


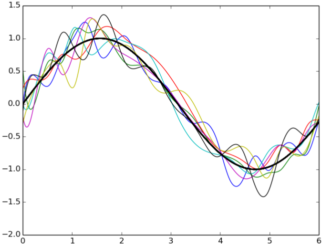
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



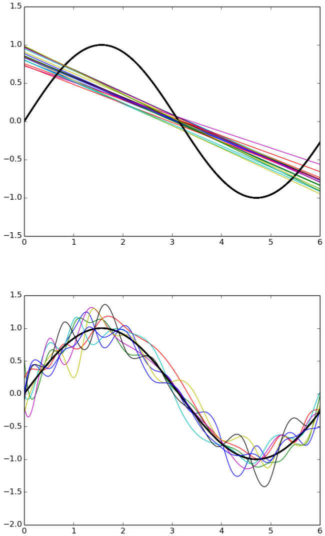


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)

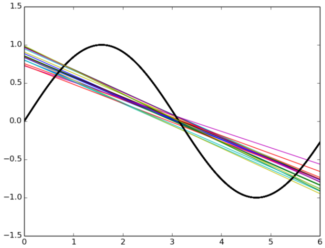


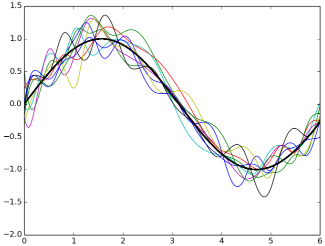


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)

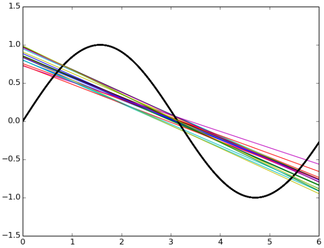


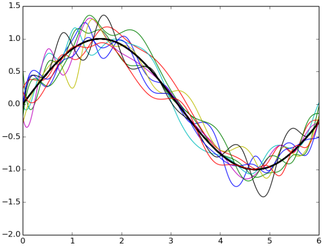
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



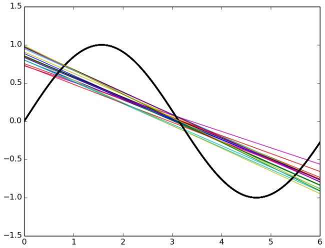


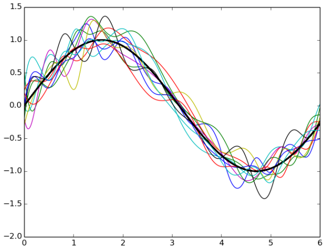
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



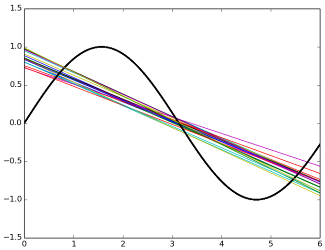


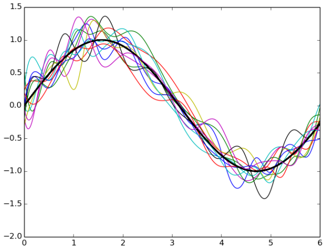
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



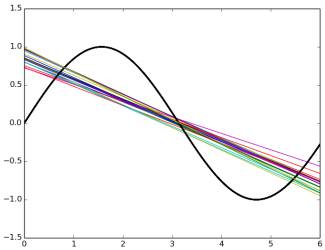


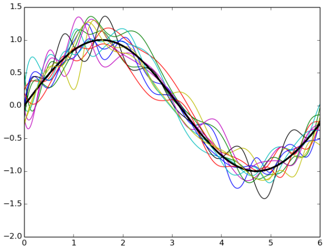
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



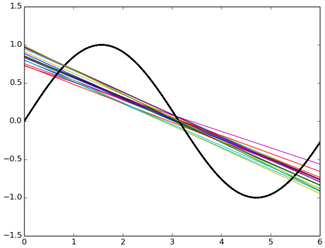


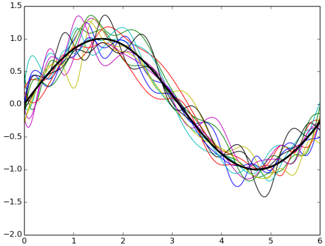
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



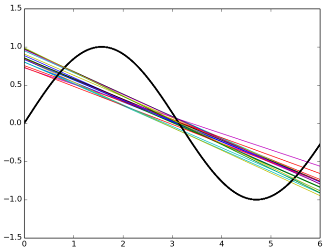


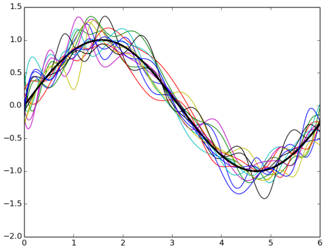
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



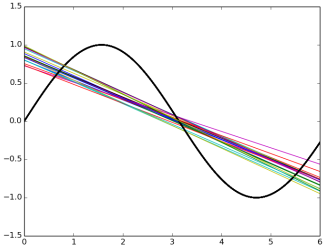


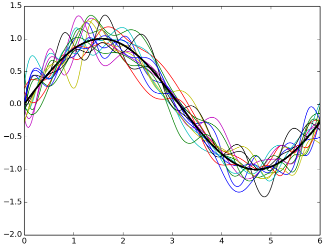
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



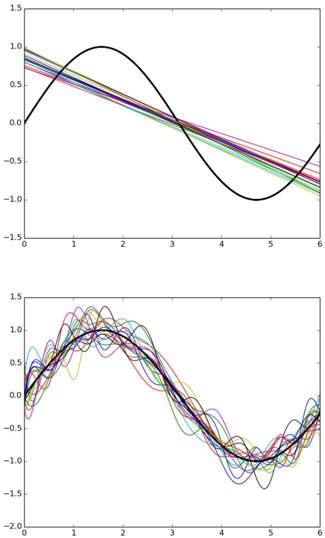


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)

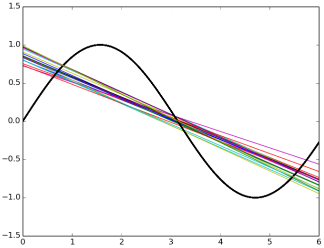


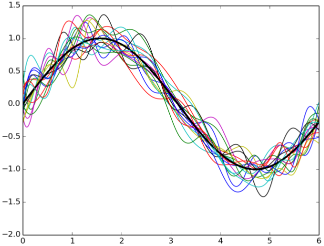


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)

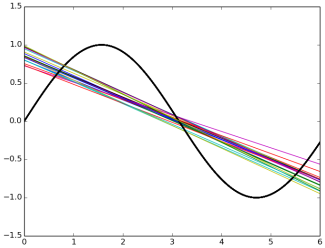


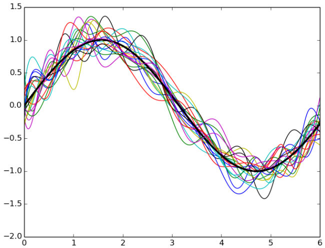
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)



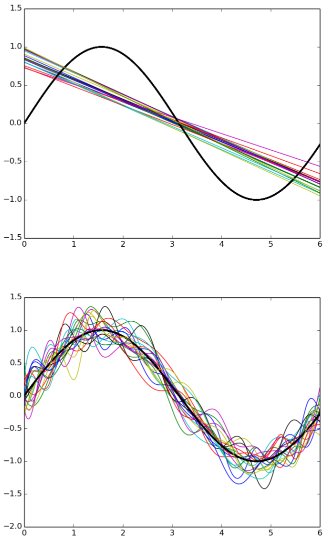


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)

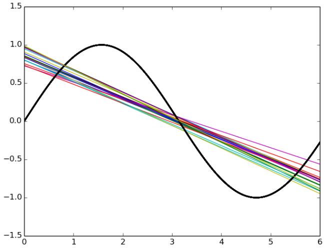


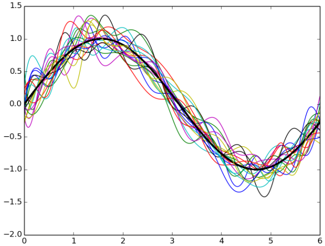


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)

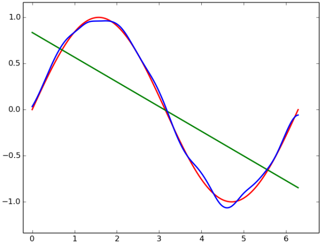


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)





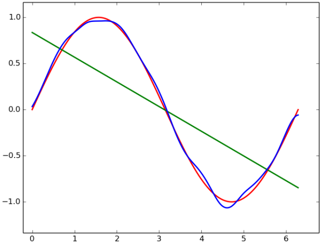
- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)
- On the other hand, complex models trained on different samples of the data are very different from each other (high variance)



Green Line: Average value of $\hat{f}(x)$ for the simple model
Blue Curve: Average value of $\hat{f}(x)$ for the complex model
Red Curve: True model ($f(x)$)

- Let $f(x)$ be the true model (sinusoidal in this case) and $\hat{f}(x)$ be our estimate of the model (simple or complex, in this case) then,

$$\text{Bias}\ (\hat{f}(x)) = E[\hat{f}(x)] - f(x)$$

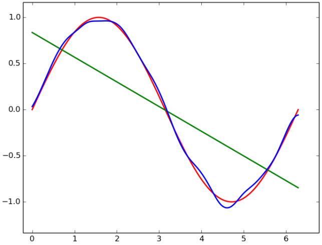


Green Line: Average value of $\hat{f}(x)$ for the simple model
Blue Curve: Average value of $\hat{f}(x)$ for the complex model
Red Curve: True model $(f(x))$

- Let $f(x)$ be the true model (sinusoidal in this case) and $\hat{f}(x)$ be our estimate of the model (simple or complex, in this case) then,

$$\text{Bias }(\hat{f}(x)) = E[\hat{f}(x)] - f(x)$$

- $E[\hat{f}(x)]$ is the average (or expected) value of the model

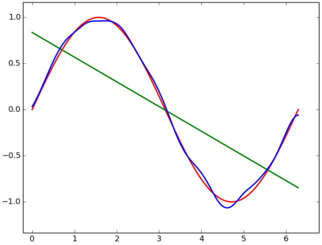


<u>Green Line</u>: Average value of $\hat{f}(x)$ for the simple model
<u>Blue Curve</u>: Average value of $\hat{f}(x)$ for the complex model
<u>Red Curve</u>: True model ($f(x)$)

- Let $f(x)$ be the true model (sinusoidal in this case) and $\hat{f}(x)$ be our estimate of the model (simple or complex, in this case) then,

$$\text{Bias}\ (\hat{f}(x)) = E[\hat{f}(x)] - f(x)$$

- $E[\hat{f}(x)]$ is the average (or expected) value of the model
- We can see that for the simple model the average value (blue line) is very far from the true value $f(x)$ (sinusoidal function)

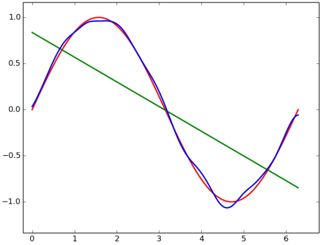


Green Line: Average value of $\hat{f}(x)$ for the simple model
Blue Curve: Average value of $\hat{f}(x)$ for the complex model
Red Curve: True model ($f(x)$)

- Let $f(x)$ be the true model (sinusoidal in this case) and $\hat{f}(x)$ be our estimate of the model (simple or complex, in this case) then,

$$\text{Bias } (\hat{f}(x)) = E[\hat{f}(x)] - f(x)$$

- $E[\hat{f}(x)]$ is the average (or expected) value of the model
- We can see that for the simple model the average value (blue line) is very far from the true value $f(x)$ (sinusoidal function)
- Mathematically, this means that the simple model has a high bias

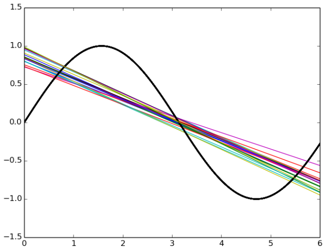


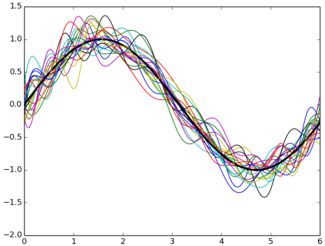
Green Line: Average value of $\hat{f}(x)$ for the simple model
Blue Curve: Average value of $\hat{f}(x)$ for the complex model
Red Curve: True model ($f(x)$)

- Let $f(x)$ be the true model (sinusoidal in this case) and $\hat{f}(x)$ be our estimate of the model (simple or complex, in this case) then,

$$\text{Bias } (\hat{f}(x)) = E[\hat{f}(x)] - f(x)$$

- $E[\hat{f}(x)]$ is the average (or expected) value of the model
- We can see that for the simple model the average value (blue line) is very far from the true value $f(x)$ (sinusoidal function)
- Mathematically, this means that the simple model has a high bias
- On the other hand, the complex model has a low bias

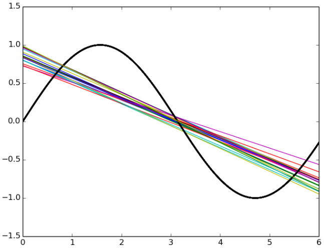


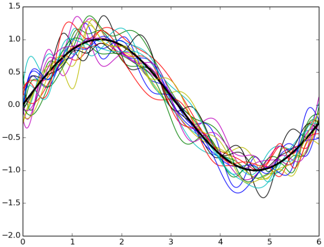


- We now define,

$$\text{Variance } (\hat{f}(x)) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

(Standard definition from statistics)



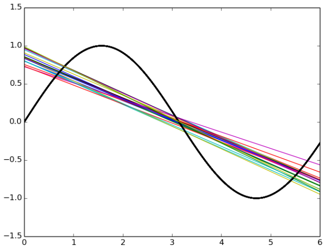


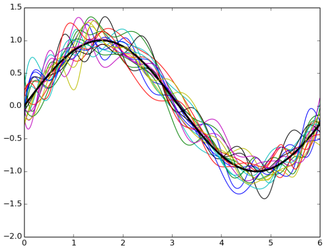
- We now define,

$$\text{Variance } (\hat{f}(x)) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

(Standard definition from statistics)

- Roughly speaking it tells us how much the different $f(x)$'s (trained on different samples of the data) differ from each other





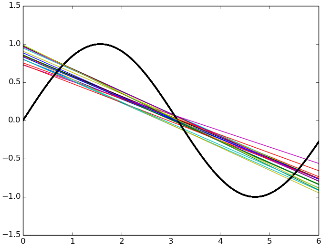
- We now define,

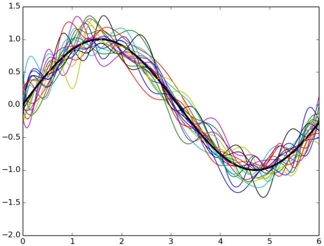
$$\text{Variance } (\hat{f}(x)) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

(Standard definition from statistics)

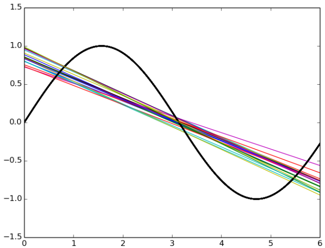
- Roughly speaking it tells us how much the different $f(x)$'s (trained on different samples of the data) differ from each other

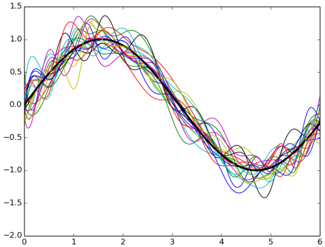
- It is clear that the simple model has a low variance whereas the complex model has a high variance



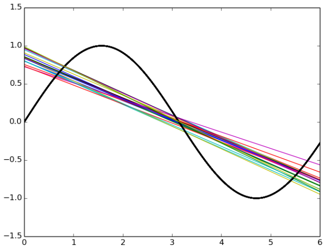


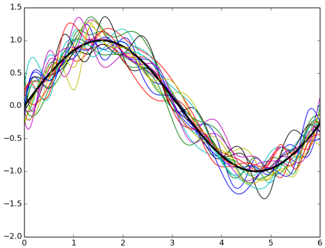
- In summary (informally)



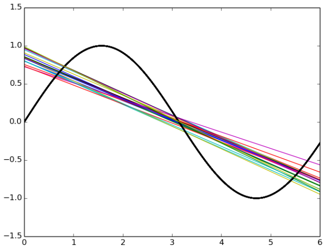


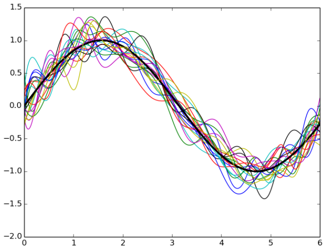
- In summary (informally)
- Simple model: high bias, low variance



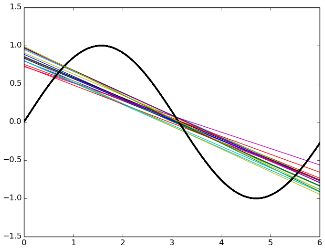


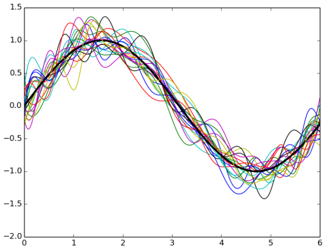
- In summary (informally)
- Simple model: high bias, low variance
- Complex model: low bias, high variance





- In summary (informally)
- Simple model: high bias, low variance
- Complex model: low bias, high variance
- There is always a trade-off between the bias and variance





- In summary (informally)
- Simple model: high bias, low variance
- Complex model: low bias, high variance
- There is always a trade-off between the bias and variance
- Both bias and variance contribute to the mean square error. Let us see how,

- Consider a new point (x,y) which was not seen during training

- Consider a new point (x,y) which was not seen during training
- If we use the model $\hat{f}(x)$ to predict the value of y then the mean square error is given by

$$E[(y - \hat{f}(x))^2]$$

(average square error in predicting y for many such unseen points)

- Consider a new point (x,y) which was not seen during training
- If we use the model $\hat{f}(x)$ to predict the value of y then the mean square error is given by

$$E[(y - \hat{f}(x))^2]$$

(average square error in predicting y for many such unseen points)

- We can show that

$$\begin{aligned} E[(y - \hat{f}(x))^2] = & \; Bias^2 \\ & + Variance \\ & + \sigma^2 \text{ (irreducible error)} \end{aligned}$$

- We can show that

$$E[(y - \hat{f}(x))^2] = Bias^2 + Variance + \sigma^2 \text{ (irreducible error)}$$

- See proof here

- Consider a new point (x,y) which was not seen during training
- If we use the model $\hat{f}(x)$ to predict the value of y then the mean square error is given by

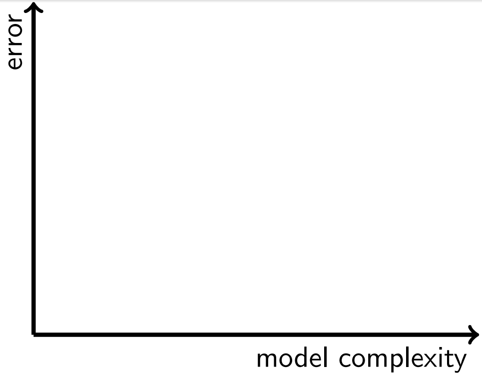
$$E[(y - \hat{f}(x))^2]$$

(average square error in predicting y for many such unseen points)

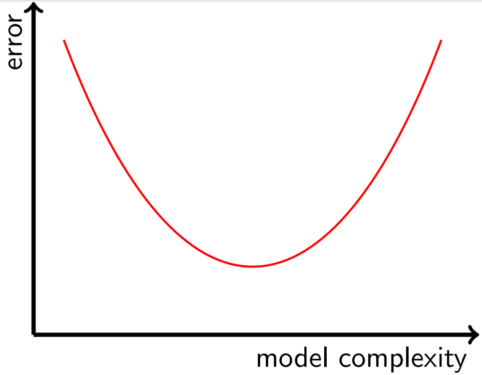
- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^{n}$

- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^{n}$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training

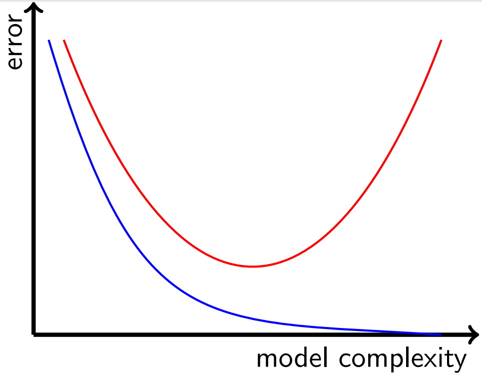
- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^{n}$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:
  $train_{err}$ (say, mean square error)
  $test_{err}$ (say, mean square error)



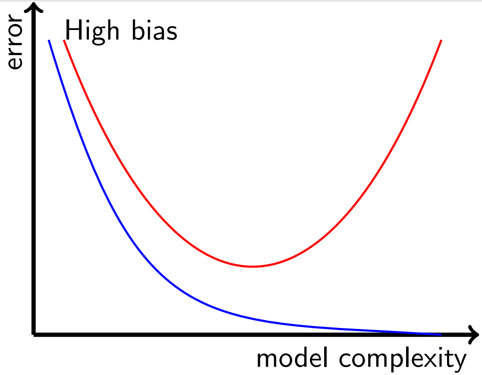
- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^{n}$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:
  $train_{err}$ (say, mean square error)
  $test_{err}$ (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



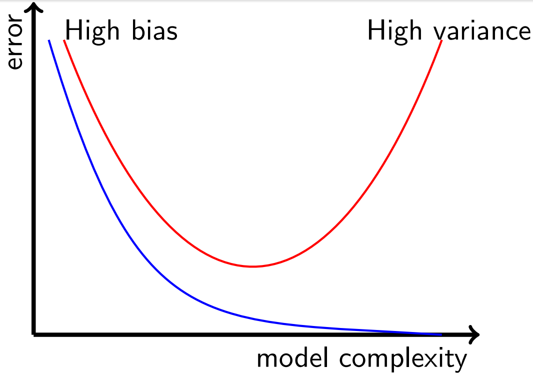
- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^{n}$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:
  $train_{err}$ (say, mean square error)
  $test_{err}$ (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



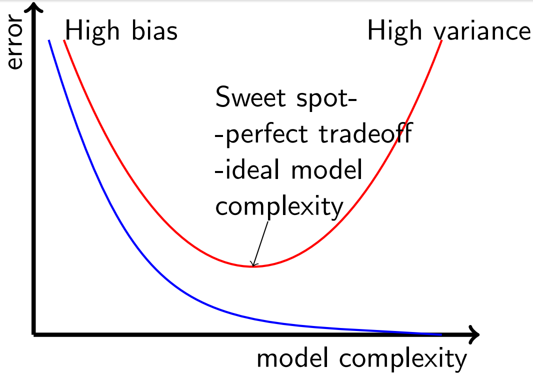
- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^{n}$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:
  $train_{err}$ (say, mean square error)
  $test_{err}$ (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



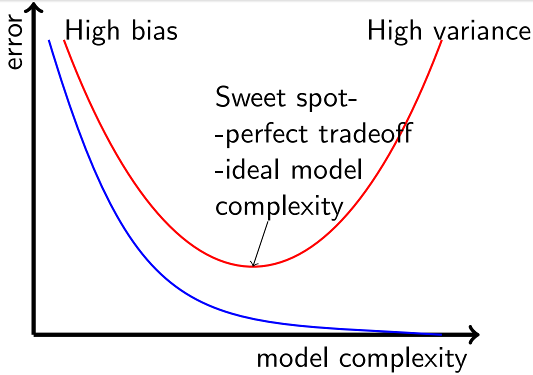
- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^{n}$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:
  $train_{err}$ (say, mean square error)
  $test_{err}$ (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:
  $train_{err}$ (say, mean square error)
  $test_{err}$ (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^{n}$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:
  $train_{err}$ (say, mean square error)
  $test_{err}$ (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



$$E[(y - \hat{f}(x))^2] = Bias^2 + Variance + \sigma^2 \text{ (irreducible error)}$$

- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:
  $train_{err}$ (say, mean square error)
  $test_{err}$ (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure

Intuitions developed so far

- Let there be n training points and m test (validation) points

$$train_{err} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

$$test_{err} = \frac{1}{m}\sum_{i=n+1}^{n+m}(y_i - \hat{f}(x_i))$$

Intuitions developed so far

- Let there be n training points and m test (validation) points

$$train_{err} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

$$test_{err} = \frac{1}{m}\sum_{i=n+1}^{n+m}(y_i - \hat{f}(x_i))$$

- As the model complexity increases $train_{err}$ becomes overly optimistic and gives us a wrong picture of how close $\hat{f}$ is to f

Intuitions developed so far

- Let there be n training points and m test (validation) points

$$train_{err} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

$$test_{err} = \frac{1}{m}\sum_{i=n+1}^{n+m}(y_i - \hat{f}(x_i))$$

- As the model complexity increases $train_{err}$ becomes overly optimistic and gives us a wrong picture of how close $\hat{f}$ is to f

- The validation error gives the real picture of how close $\hat{f}$ is to f

Intuitions developed so far

- Let there be n training points and m test (validation) points

$$train_{err} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

$$test_{err} = \frac{1}{m}\sum_{i=n+1}^{n+m}(y_i - \hat{f}(x_i))$$

- As the model complexity increases $train_{err}$ becomes overly optimistic and gives us a wrong picture of how close $\hat{f}$ is to f
- The validation error gives the real picture of how close $\hat{f}$ is to f
- We will concretize this intuition mathematically now and eventually show how to account for the optimism in the training error

- Let D=$\{x_i, y_i\}_{i=1}^{m+n}$, then for any point (x,y) we have,

$$y_i = f(x_i) + \varepsilon_i$$

- Let $D=\{x_i, y_i\}_{i=1}^{m+n}$, then for any point (x,y) we have,

  $$y_i = f(x_i) + \varepsilon_i$$

- which means that y_i is related to x_i by some true function f but there is also some noise ε in the relation

- Let $D=\{x_i, y_i\}_{i=1}^{m+n}$, then for any point (x,y) we have,

  $$y_i = f(x_i) + \varepsilon_i$$

- which means that y_i is related to x_i by some true function f but there is also some noise ε in the relation

- For simplicity, we assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

- Let D$=\{x_i, y_i\}_{i=1}^{m+n}$, then for any point (x,y) we have,

  $$y_i = f(x_i) + \varepsilon_i$$

- which means that y_i is related to x_i by some true function f but there is also some noise ε in the relation
- For simplicity, we assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and of course we do not know f

- Let $D = \{x_i, y_i\}_{i=1}^{m+n}$, then for any point (x,y) we have,

  $y_i = f(x_i) + \varepsilon_i$

- which means that y_i is related to x_i by some true function f but there is also some noise ε in the relation
- For simplicity, we assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and of course we do not know f
- Further we use $\hat{f}$ to approximate f and estimate the parameters using T $\subset$ D such that

  $y_i = \hat{f}(x_i)$

- Let D=$\{x_i, y_i\}_{i=1}^{m+n}$, then for any point (x,y) we have,

  $y_i = f(x_i) + \varepsilon_i$

- which means that y_i is related to x_i by some true function f but there is also some noise ε in the relation

- For simplicity, we assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and of course we do not know f

- Further we use $\hat{f}$ to approximate f and estimate the parameters using T $\subset$ D such that

  $y_i = \hat{f}(x_i)$

- We are interested in knowing $E[(\hat{f}(x_i) - f(x_i))^2]$

- Let $D=\{x_i, y_i\}_{i=1}^{m+n}$, then for any point (x,y) we have,

  $y_i = f(x_i) + \varepsilon_i$

- which means that y_i is related to x_i by some true function f but there is also some noise ε in the relation
- For simplicity, we assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and of course we do not know f
- Further we use $\hat{f}$ to approximate f and estimate the parameters using T $\subset$ D such that

  $y_i = \hat{f}(x_i)$

- We are interested in knowing $E[(\hat{f}(x_i) - f(x_i))^2]$ but we cannot estimate this directly because we do not know f

- Let D$=\{x_i, y_i\}_{i=1}^{m+n}$, then for any point (x,y) we have,

  $y_i = f(x_i) + \varepsilon_i$

- which means that y_i is related to x_i by some true function f but there is also some noise ε in the relation
- For simplicity, we assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and of course we do not know f
- Further we use $\hat{f}$ to approximate f and estimate the parameters using T $\subset$ D such that
  $y_i = \hat{f}(x_i)$
- We are interested in knowing $E[(\hat{f}(x_i) - f(x_i))^2]$
  but we cannot estimate this directly because we do not know f
- We will see how to estimate this empirically using the observation y_i & prediction $\hat{y}_i$

$$
\begin{aligned}
&E[(\hat{y}_i - y_i)^2] \\
&\quad = E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i)
\end{aligned}
$$

$$
\begin{aligned}
&E[(\hat{y}_i - y_i)^2] \\
&\quad = E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i) \\
&\quad = E[(\hat{f}(x_i) - f(x_i))^2 - 2\varepsilon_i(\hat{f}(x_i) - f(x_i)) + \varepsilon_i^2]
\end{aligned}
$$

$$E[(\hat{y}_i - y_i)^2]$$

$$= E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i)$$

$$= E[(\hat{f}(x_i) - f(x_i))^2 - 2\varepsilon_i(\hat{f}(x_i) - f(x_i)) + \varepsilon_i^2]$$

$$= E[(\hat{f}(x_i) - f(x_i))^2] - 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))] + E[\varepsilon_i^2]$$

$$
\begin{aligned}
&E[(\hat{y}_i - y_i)^2] \\
&\quad = E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i) \\
&\quad = E[(\hat{f}(x_i) - f(x_i))^2 - 2\varepsilon_i(\hat{f}(x_i) - f(x_i)) + \varepsilon_i^2] \\
&\quad = E[(\hat{f}(x_i) - f(x_i))^2] - 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))] + E[\varepsilon_i^2] \\
&\quad = E[(\hat{f}(x_i) - f(x_i))^2] - 2E[(y_i - f(x_i))(\hat{f}(x_i) - f(x_i))] + E[\varepsilon_i^2] \quad (\varepsilon_i = y_i - f(x_i))
\end{aligned}
$$

$$E[(\hat{y}_i - y_i)^2]$$

$$= E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i)$$

$$= E[(\hat{f}(x_i) - f(x_i))^2 - 2\varepsilon_i(\hat{f}(x_i) - f(x_i)) + \varepsilon_i^2]$$

$$= E[(\hat{f}(x_i) - f(x_i))^2] - 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))] + E[\varepsilon_i^2]$$

$$\cancel{= E[(\hat{f}(x_i) - f(x_i))^2] - 2E[(y_i - f(x_i))(\hat{f}(x_i) - f(x_i))] + E[\varepsilon_i^2] \quad (\varepsilon_i = y_i - f(x_i))}$$

$$E[(\hat{y}_i - y_i)^2]$$

$$= E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i)$$

$$= E[(\hat{f}(x_i) - f(x_i))^2 - 2\varepsilon_i(\hat{f}(x_i) - f(x_i)) + \varepsilon_i^2]$$

$$= E[(\hat{f}(x_i) - f(x_i))^2] - 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))] + E[\varepsilon_i^2]$$

$$\cancel{= E[(\hat{f}(x_i) - f(x_i))^2] - 2E[(y_i - f(x_i))(\hat{f}(x_i) - f(x_i))] + E[\varepsilon_i^2] \quad (\varepsilon_i = y_i - f(x_i))}$$

$$\therefore \; E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\,\varepsilon_i(\hat{f}(x_i) - f(x_i))\,]$$

We will take a small detour to understand how to empirically estimate an Expectation and then return to our derivation

- Suppose we have observed the goals scored(z) in k matches as z_1=2, z_2=1, z_3=0, ... z_k=2

- Suppose we have observed the goals scored(z) in k matches as z_1=2, z_2=1, z_3=0, ... z_k=2
- Now we can empirically estimate $E[z]$ i.e. the expected number of goals scored as

$$E[z] = \frac{1}{k}\sum_{i=1}^{k} z_i$$

- Suppose we have observed the goals scored(z) in k matches as $z_1{=}2$, $z_2{=}1$, $z_3{=}0$, ... $z_k{=}2$

- Now we can empirically estimate $E[z]$ i.e. the expected number of goals scored as

$$E[z] = \frac{1}{k}\sum_{i=1}^{k} z_i$$

- Analogy with our derivation: We have a certain number of observations y_i & predictions $\hat{y}_i$ using which we can estimate

$$E[(\hat{y}_i - y_i)^2] =$$

- Suppose we have observed the goals scored(z) in k matches as $z_1{=}2,\ z_2{=}1,\ z_3{=}0,\ \ldots\ z_k{=}2$

- Now we can empirically estimate $E[z]$ i.e. the expected number of goals scored as

$$E[z] = \frac{1}{k}\sum_{i=1}^{k} z_i$$

- Analogy with our derivation: We have a certain number of observations y_i & predictions $\hat{y}_i$ using which we can estimate

$$E[(\hat{y}_i - y_i)^2] = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2$$

... returning back to our derivation

$$E[(\hat{f}(x_i) - f(x_i))^2] \; = \; E[(\hat{y}_i - y_i)^2] \; - \; E[\varepsilon_i^2] \; + \; 2E[\; \varepsilon_i(\hat{f}(x_i) - f(x_i)) \;]$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]$$

- We can empirically evaluate R.H.S using training observations or test observations

Case 1: Using test observations

$$E[(\hat{f}(x_i) - f(x_i))^2] \;=\; E[(\hat{y}_i - y_i)^2] \;-\; E[\varepsilon_i^2] \;+\; 2E[\;\varepsilon_i(\hat{f}(x_i) - f(x_i))\;]$$

- We can empirically evaluate R.H.S using training observations or test observations

Case 1: Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\textit{true error}}$$

$$E[(\hat{f}(x_i) - f(x_i))^2] \; = \; E[(\hat{y}_i - y_i)^2] \; - \; E[\varepsilon_i^2] \; + \; 2E[\; \varepsilon_i(\hat{f}(x_i) - f(x_i)) \;]$$

- We can empirically evaluate R.H.S using training observations or test observations

Case 1: Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error} \quad = \quad \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ \ error} \quad -$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]$$

- We can empirically evaluate R.H.S using training observations or test observations

Case 1: Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error} = \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant}$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]$$

- We can empirically evaluate R.H.S using training observations or test observations

Case 1: Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{\text{small constant}} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{\simeq \text{covariance}\,(y,\hat{f}(x))}$$

$$E[(\hat{f}(x_i) - f(x_i))^2] \;=\; E[(\hat{y}_i - y_i)^2] \;-\; E[\varepsilon_i^2] \;+\; 2E[\;\varepsilon_i(\hat{f}(x_i) - f(x_i))\;]$$

- We can empirically evaluate R.H.S using training observations or test observations

Case 1: Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \;=\; \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} \;-\; \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{\text{small constant}} \;+\; \underbrace{2E[\;\varepsilon_i(\hat{f}(x_i) - f(x_i))\;]}_{=\,\text{covariance}\,(\varepsilon_i,\hat{f}(x_i)-f(x_i))}$$

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\,(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\,(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

- None of the test observations participated in the estimation of $\hat{f}(x)$[the parameters of $\hat{f}(\mathrm{x})$ were estimated only using training data]

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\,(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

- None of the test observations participated in the estimation of $\hat{f}(x)$[the parameters of $\hat{f}(x)$ were estimated only using training data]
  $\therefore\ \varepsilon \perp (\hat{f}(x_i) - f(x_i))$

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\ (\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

- None of the test observations participated in the estimation of $\hat{f}(x)$[the parameters of $\hat{f}(x)$ were estimated only using training data]
  $\therefore\ \varepsilon \perp (\hat{f}(x_i) - f(x_i))$
  $\therefore\ E[\varepsilon_i\ .\ (\hat{f}(x_i) - f(x_i))]$

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m} \varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\,(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

- None of the test observations participated in the estimation of $\hat{f}(x)$[the parameters of $\hat{f}(x)$ were estimated only using training data]
  $\therefore\ \varepsilon \perp (\hat{f}(x_i) - f(x_i))$
  $\therefore\ E[\varepsilon_i\ .\ (\hat{f}(x_i) - f(x_i))] = E[\varepsilon_i]\ .\ E[\hat{f}(x_i) - f(x_i))]$

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\,(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

- None of the test observations participated in the estimation of $\hat{f}(x)$[the parameters of $\hat{f}(x)$ were estimated only using training data]
  $\therefore\ \varepsilon \perp (\hat{f}(x_i) - f(x_i))$
  $\therefore\ E[\varepsilon_i\ .\ (\hat{f}(x_i) - f(x_i))] = E[\varepsilon_i]\ .\ E[\hat{f}(x_i) - f(x_i))] = 0\ .\ E[\hat{f}(x_i) - f(x_i))]$

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\ (\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

- None of the test observations participated in the estimation of $\hat{f}(x)$[the parameters of $\hat{f}(\mathrm{x})$ were estimated only using training data]

$\therefore\ \ \varepsilon \perp (\hat{f}(x_i) - f(x_i))$

$\therefore\ \ E[\varepsilon_i\ .\ (\hat{f}(x_i) - f(x_i))] = E[\varepsilon_i]\ .\ E[\hat{f}(x_i) - f(x_i))] = 0\ .\ E[\hat{f}(x_i) - f(x_i))] = 0$

$\therefore\ \ \text{true error} = \text{empirical test error} + \text{small constant}$

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=covariance\,(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

- None of the test observations participated in the estimation of $\hat{f}(x)$[the parameters of $\hat{f}(x)$ were estimated only using training data]
  $\therefore\ \varepsilon \perp (\hat{f}(x_i) - f(x_i))$
  $\therefore\ E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] = E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i))] = 0 \cdot E[\hat{f}(x_i) - f(x_i))] = 0$
  $\therefore$ true error = empirical test error + small constant

- Hence, we should always use a validation set(independent of the training set) to estimate the error

Case 2: Using training observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{= covariance\ (\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

Case 2: Using training observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error} = \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\ (\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

Now, $\varepsilon \not\perp \hat{f}(x)$ because ε was used for estimating the parameters of $\hat{f}(x)$

Case 2: Using training observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\textit{true error}} = \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{\textit{empirical estimation of error}} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{\textit{small constant}} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ \textit{covariance}\ (\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

Now, $\varepsilon \not\perp \hat{f}(x)$ because ε was used for estimating the parameters of $\hat{f}(x)$

$\therefore\ E[\varepsilon_i\ .\ (\hat{f}(x_i) - f(x_i))]$

Case 2: Using training observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\ (\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

Now, $\varepsilon \not\perp \hat{f}(x)$ because ε was used for estimating the parameters of $\hat{f}(x)$

$$\therefore\ E[\varepsilon_i\ .\ (\hat{f}(x_i) - f(x_i))] \neq E[\varepsilon_i]\ .\ E[\hat{f}(x_i) - f(x_i))]$$

Case 2: Using training observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error}$$

$$= \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\ (\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

Now, $\varepsilon \not\perp \hat{f}(\mathrm{x})$ because ε was used for estimating the parameters of $\hat{f}(\mathrm{x})$

$$\therefore\ E[\varepsilon_i\ .\ (\hat{f}(x_i) - f(x_i))] \neq E[\varepsilon_i]\ .\ E[\hat{f}(x_i) - f(x_i))] \neq\ 0$$

Hence, the empirical train error is smaller than the true error and does not give a true picture of the error

Case 2: Using training observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{true\ error} = \underbrace{\frac{1}{m}\sum_{i=1}^{n+m}(\hat{y}_i - y_i)^2}_{empirical\ estimation\ of\ error} - \underbrace{\frac{1}{m}\sum_{i=n+1}^{n+m}\varepsilon_i^2}_{small\ constant} + \underbrace{2E[\ \varepsilon_i(\hat{f}(x_i) - f(x_i))\ ]}_{=\ covariance\ (\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

Now, $\varepsilon \not\perp \hat{f}(x)$ because ε was used for estimating the parameters of $\hat{f}(x)$

$$\therefore\ E[\varepsilon_i\ .\ (\hat{f}(x_i) - f(x_i))] \neq E[\varepsilon_i]\ .\ E[\hat{f}(x_i) - f(x_i))] \neq\ 0$$

Hence, the empirical train error is smaller than the true error and does not give a true picture of the error

But how is this related to model complexity? Let us see

Using Stein's Lemma (and some trickery) we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$$

Using Stein's Lemma (and some trickery) we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$$

- When will $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ be high? When a small change in the observation causes a large change in the estimation($\hat{f}$)

Using Stein's Lemma (and some trickery) we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$$

- When will $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ be high? When a small change in the observation causes a large change in the estimation($\hat{f}$)
- Can you link this to model complexity?

Using Stein's Lemma (and some trickery) we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$$

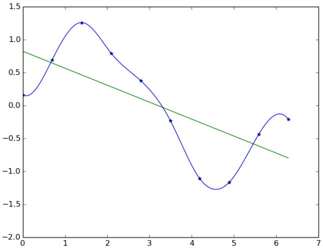
- When will $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ be high? When a small change in the observation causes a large change in the estimation($\hat{f}$)
- Can you link this to model complexity?
- Yes, indeed a complex model will be more sensitive to changes in observations whereas a simple model will be less sensitive to changes in observations

Using Stein's Lemma (and some trickery) we can show that

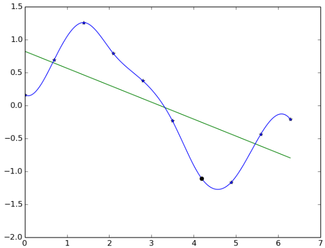
$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$$

- When will $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ be high? When a small change in the observation causes a large change in the estimation($\hat{f}$)
- Can you link this to model complexity?
- Yes, indeed a complex model will be more sensitive to changes in observations whereas a simple model will be less sensitive to changes in observations
- Hence, we can say that
  true error $=$ empirical train error $+$ small constant $+ \Omega$(model complexity)

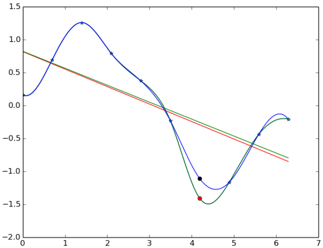
- Let us verify that indeed a complex model is more sensitive to minor changes in the data



- Let us verify that indeed a complex model is more sensitive to minor changes in the data
- We have fitted a simple and complex model for some given data



- Let us verify that indeed a complex model is more sensitive to minor changes in the data
- We have fitted a simple and complex model for some given data
- We now change one of these data points



- Let us verify that indeed a complex model is more sensitive to minor changes in the data
- We have fitted a simple and complex model for some given data
- We now change one of these data points
- The simple model does not change much as compared to the complex model

- Hence while training, instead of minimizing the training error $\mathscr{L}_{train}(\theta)$ we should minimize

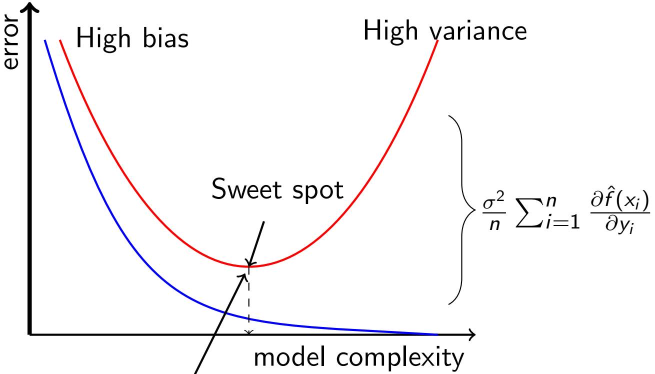
$$\min_{w.r.t\ \theta} \mathscr{L}_{train}(\theta) + \Omega(\theta) = \mathscr{L}(\theta)$$

- Hence while training, instead of minimizing the training error $\mathscr{L}_{train}(\theta)$ we should minimize
$$\min_{w.r.t\ \theta} \mathscr{L}_{train}(\theta) + \Omega(\theta) = \mathscr{L}(\theta)$$
- Where $\Omega(\theta)$ would be high for complex models and small for simple models

- Hence while training, instead of minimizing the training error $\mathscr{L}_{train}(\theta)$ we should minimize
$$\min_{w.r.t\ \theta} \mathscr{L}_{train}(\theta) + \Omega(\theta) = \mathscr{L}(\theta)$$
- Where $\Omega(\theta)$ would be high for complex models and small for simple models
- $\Omega(\theta)$ acts as an approximate for $\frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$

- Hence while training, instead of minimizing the training error $\mathscr{L}_{train}(\theta)$ we should minimize
$$\min_{w.r.t\ \theta} \mathscr{L}_{train}(\theta) + \Omega(\theta) = \mathscr{L}(\theta)$$
- Where $\Omega(\theta)$ would be high for complex models and small for simple models
- $\Omega(\theta)$ acts as an approximate for $\frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$
- This is the basis for all regularization methods

- Hence while training, instead of minimizing the training error $\mathscr{L}_{train}(\theta)$ we should minimize
  $\min_{w.r.t\ \theta} \mathscr{L}_{train}(\theta) + \Omega(\theta) = \mathscr{L}(\theta)$
- Where $\Omega(\theta)$ would be high for complex models and small for simple models
- $\Omega(\theta)$ acts as an approximate for $\frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{\partial \hat{f}(x_i)}{\partial y_i}$
- This is the basis for all regularization methods
- We can show that L_1 regularization, L_2 regularization, early stopping and injecting noise in input are all instances of this form of regularization.



$\Omega(\theta)$ should ensure that model has reasonable complexity

- Why do we care about this bias variance tradeoff and model complexity?

- Why do we care about this bias variance tradeoff and model complexity?

- Why do we care about this bias variance tradeoff and model complexity?

- Deep Neural networks are highly complex models.

- Why do we care about this bias variance tradeoff and model complexity?

- Deep Neural networks are highly complex models.
- Many parameters, many non-linearties.

- Why do we care about this bias variance tradeoff and model complexity?

- Deep Neural networks are highly complex models.
- Many parameters, many non-linearties.
- It is easy for them to overfit and drive training error to 0.

- Why do we care about this bias variance tradeoff and model complexity?

- Deep Neural networks are highly complex models.
- Many parameters, many non-linearties.
- It is easy for them to overfit and drive training error to 0.
- Hence we need some form of regularization.

Different forms of regularization

- $L2$ regularization

Different forms of regularization

- $L2$ regularization
- Dataset augmentation

Different forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying

Different forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs

Different forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs

Different forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping

Different forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods

Different forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

Different forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

- For L2 regularization we have,

$$\widetilde{J}(w) = J(w) + \frac{\alpha}{2}\|w\|^2$$

- For L2 regularization we have,

$$\widetilde{J}(w) = J(w) + \frac{\alpha}{2}\|w\|^2$$

- For SGD (or its variants), we are interested in

$$\nabla \widetilde{J}(w) = \nabla J(w) + \alpha w$$

- For L2 regularization we have,

$$\widetilde{J}(w) = J(w) + \frac{\alpha}{2}\|w\|^2$$

- For SGD (or its variants), we are interested in

$$\nabla \widetilde{J}(w) = \nabla J(w) + \alpha w$$

- Update rule:

$$w_{t+1} = w_t - \eta \nabla J(w_t) - \eta \alpha w_t$$

- For L2 regularization we have,

$$\widetilde{J}(w) = J(w) + \frac{\alpha}{2}\|w\|^2$$

- For SGD (or its variants), we are interested in

$$\nabla \widetilde{J}(w) = \nabla J(w) + \alpha w$$

- Update rule:

$$w_{t+1} = w_t - \eta \nabla J(w_t) - \eta \alpha w_t$$

- Requires a very small modification to the code

- For L2 regularization we have,

$$\widetilde{J}(w) = J(w) + \frac{\alpha}{2}\|w\|^2$$

- For SGD (or its variants), we are interested in

$$\nabla \widetilde{J}(w) = \nabla J(w) + \alpha w$$

- Update rule:

$$w_{t+1} = w_t - \eta \nabla J(w_t) - \eta \alpha w_t$$

- Requires a very small modification to the code
- Let us see the geometric interpretation of this

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$J(w) = J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H(w - w^*)$$

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$
\begin{aligned}
J(w) &= J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H(w - w^*) \\
&= J(w^*) + \frac{1}{2}(w - w^*)^T H(w - w^*) \qquad (\because \nabla J(w^*) = 0\ )
\end{aligned}
$$

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$
\begin{aligned}
J(w) &= J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \\
&= J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \qquad (\because \nabla J(w^*) = 0\ )
\end{aligned}
$$

$$
\nabla J(w) = \nabla J(w^*) + H(w - w^*)
$$

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$
\begin{aligned}
J(w) &= J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \\
&= J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \qquad (\because \nabla J(w^*) = 0\ ) \\
\nabla J(w) &= \nabla J(w^*) + H(w - w^*) \\
&= H(w - w^*)
\end{aligned}
$$

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$
\begin{aligned}
J(w) &= J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \\
&= J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \qquad (\because \nabla J(w^*) = 0\ ) \\
\nabla J(w) &= \nabla J(w^*) + H(w - w^*) \\
&= H(w - w^*)
\end{aligned}
$$

- Now,

$$\nabla \widetilde{J}(w) = \nabla J(w) + \alpha w$$

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$\begin{aligned}
J(w) &= J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H(w - w^*) \\
&= J(w^*) + \frac{1}{2}(w - w^*)^T H(w - w^*) \qquad (\because \nabla J(w^*) = 0\ ) \\
\nabla J(w) &= \nabla J(w^*) + H(w - w^*) \\
&= H(w - w^*)
\end{aligned}$$

- Now,

$$\begin{aligned}
\nabla \widetilde{J}(w) &= \nabla J(w) + \alpha w \\
&= H(w - w^*) + \alpha w
\end{aligned}$$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$H(\widetilde{w} - w^*) + \alpha\widetilde{w} = 0$$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$H(\widetilde{w} - w^*) + \alpha \widetilde{w} = 0$$
$$(H + \alpha \mathbb{I})\widetilde{w} = Hw^*$$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$\begin{aligned} H(\widetilde{w} - w^*) + \alpha \widetilde{w} &= 0 \\ (H + \alpha \mathbb{I})\widetilde{w} &= H w^* \\ \widetilde{w} &= (H + \alpha \mathbb{I})^{-1} H w^* \end{aligned}$$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$\begin{aligned} H(\widetilde{w} - w^*) + \alpha \widetilde{w} &= 0 \\ (H + \alpha \mathbb{I})\widetilde{w} &= Hw^* \\ \widetilde{w} &= (H + \alpha \mathbb{I})^{-1} Hw^* \end{aligned}$$

- Notice that if $\alpha \rightarrow 0$ then $\widetilde{w} \rightarrow w^*$ [no regularization]

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$\begin{aligned} H(\widetilde{w} - w^*) + \alpha \widetilde{w} &= 0 \\ (H + \alpha \mathbb{I})\widetilde{w} &= H w^* \\ \widetilde{w} &= (H + \alpha \mathbb{I})^{-1} H w^* \end{aligned}$$

- Notice that if $\alpha \rightarrow 0$ then $\widetilde{w} \rightarrow w^*$ [no regularization]
- But we are interested in the case when $\alpha \neq 0$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$\begin{aligned} H(\widetilde{w} - w^*) + \alpha \widetilde{w} &= 0 \\ (H + \alpha \mathbb{I})\widetilde{w} &= Hw^* \\ \widetilde{w} &= (H + \alpha \mathbb{I})^{-1} Hw^* \end{aligned}$$

- Notice that if $\alpha \to 0$ then $\widetilde{w} \to w^*$ [no regularization]
- But we are interested in the case when $\alpha \neq 0$
- Let us analyse the case when $\alpha \neq 0$

- If H is symmetric Positive Semi Definite

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\widetilde{w} = (H + \alpha\mathbb{I})^{-1}Hw^*$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\begin{aligned}\widetilde{w} &= (H + \alpha\mathbb{I})^{-1}Hw^* \\ &= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1}Q\Lambda Q^T w^*\end{aligned}$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\begin{aligned}
\widetilde{w} &= (H + \alpha\mathbb{I})^{-1}Hw^* \\
&= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1}Q\Lambda Q^T w^* \\
&= (Q\Lambda Q^T + \alpha Q\mathbb{I}Q^T)^{-1}Q\Lambda Q^T w^*
\end{aligned}$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\begin{aligned}
\widetilde{w} &= (H + \alpha\mathbb{I})^{-1} H w^* \\
&= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1} Q\Lambda Q^T w^* \\
&= (Q\Lambda Q^T + \alpha Q\mathbb{I}Q^T)^{-1} Q\Lambda Q^T w^* \\
&= [Q(\Lambda + \alpha\mathbb{I})Q^T]^{-1} Q\Lambda Q^T w^*
\end{aligned}$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\begin{aligned}
\widetilde{w} &= (H + \alpha\mathbb{I})^{-1} H w^* \\
&= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1} Q\Lambda Q^T w^* \\
&= (Q\Lambda Q^T + \alpha Q\mathbb{I}Q^T)^{-1} Q\Lambda Q^T w^* \\
&= [Q(\Lambda + \alpha\mathbb{I})Q^T]^{-1} Q\Lambda Q^T w^* \\
&= Q^{T^{-1}}(\Lambda + \alpha\mathbb{I})^{-1} Q^{-1} Q\Lambda Q^T w^*
\end{aligned}$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\begin{aligned}
\widetilde{w} &= (H + \alpha\mathbb{I})^{-1}Hw^* \\
&= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1}Q\Lambda Q^T w^* \\
&= (Q\Lambda Q^T + \alpha Q\mathbb{I}Q^T)^{-1}Q\Lambda Q^T w^* \\
&= [Q(\Lambda + \alpha\mathbb{I})Q^T]^{-1}Q\Lambda Q^T w^* \\
&= Q^{T^{-1}}(\Lambda + \alpha\mathbb{I})^{-1}Q^{-1}Q\Lambda Q^T w^* \\
&= Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^* \qquad (\because Q^{T^{-1}} = Q)
\end{aligned}$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\begin{aligned}
\widetilde{w} &= (H + \alpha\mathbb{I})^{-1}Hw^* \\
&= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1}Q\Lambda Q^T w^* \\
&= (Q\Lambda Q^T + \alpha Q\mathbb{I}Q^T)^{-1}Q\Lambda Q^T w^* \\
&= [Q(\Lambda + \alpha\mathbb{I})Q^T]^{-1}Q\Lambda Q^T w^* \\
&= Q^{T^{-1}}(\Lambda + \alpha\mathbb{I})^{-1}Q^{-1}Q\Lambda Q^T w^* \\
&= Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^* \qquad (\because Q^{T^{-1}} = Q) \\
\widetilde{w} &= QDQ^T w^*
\end{aligned}$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\begin{aligned}
\widetilde{w} &= (H + \alpha\mathbb{I})^{-1}Hw^* \\
&= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1}Q\Lambda Q^T w^* \\
&= (Q\Lambda Q^T + \alpha Q\mathbb{I}Q^T)^{-1}Q\Lambda Q^T w^* \\
&= [Q(\Lambda + \alpha\mathbb{I})Q^T]^{-1}Q\Lambda Q^T w^* \\
&= Q^{T^{-1}}(\Lambda + \alpha\mathbb{I})^{-1}Q^{-1}Q\Lambda Q^T w^* \\
&= Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^* \qquad (\because Q^{T^{-1}} = Q) \\
\widetilde{w} &= QDQ^T w^*
\end{aligned}$$

where $D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$, is a diagonal matrix which we will see in more detail soon

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

- So what is happening here?

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$

$$\begin{aligned}\widetilde{w} &= Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^* \\ &= QDQ^T w^*\end{aligned}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*

$$\widetilde{w} = Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^*$$
$$= QDQ^T w^*$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\widetilde{w} = Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha \mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1 + \alpha_1} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & & \\ & & & \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\begin{aligned}\widetilde{w} &= Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^* \\ &= QDQ^T w^*\end{aligned}$$

$$(\Lambda + \alpha \mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1 + \alpha_1} & & & \\ & \frac{1}{\lambda_2 + \alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n + \alpha_n} \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\tilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}$$

$$D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$
\begin{aligned}
\widetilde{w} &= Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^* \\
&= QDQ^T w^*
\end{aligned}
$$

$$
(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}
$$

$$
D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda
$$

$$
(\Lambda + \alpha\mathbb{I})^{-1}\Lambda = \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}
$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\widetilde{w} = Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^*$$

$$= QDQ^T w^*$$

$$(\Lambda + \alpha \mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1 + \alpha_1} & & & \\ & \frac{1}{\lambda_2 + \alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n + \alpha_n} \end{bmatrix}$$

$$D = (\Lambda + \alpha \mathbb{I})^{-1} \Lambda$$

$$(\Lambda + \alpha \mathbb{I})^{-1} \Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1 + \alpha_1} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}$$

$$D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$$

$$(\Lambda + \alpha\mathbb{I})^{-1}\Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\alpha_1} & & & \\ & \frac{\lambda_2}{\lambda_2+\alpha_2} & & \\ & & & \\ & & & \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}$$

$$D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$$

$$(\Lambda + \alpha\mathbb{I})^{-1}\Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\alpha_1} & & & \\ & \frac{\lambda_2}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}$$

$$D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$$

$$(\Lambda + \alpha\mathbb{I})^{-1}\Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\alpha_1} & & & \\ & \frac{\lambda_2}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{\lambda_n}{\lambda_n+\alpha_n} \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\begin{aligned}\widetilde{w} &= Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^* \\ &= QDQ^T w^*\end{aligned}$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}$$

$$D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$$

$$(\Lambda + \alpha\mathbb{I})^{-1}\Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\alpha_1} & & & \\ & \frac{\lambda_2}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{\lambda_n}{\lambda_n+\alpha_n} \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}$$

$$D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$$

$$(\Lambda + \alpha\mathbb{I})^{-1}\Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\alpha_1} & & & \\ & \frac{\lambda_2}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{\lambda_n}{\lambda_n+\alpha_n} \end{bmatrix}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$
- However if $\alpha = 0$ then Q rotates $Q^T w^*$ back to give w^*
- If $\alpha \neq 0$ then let us see what D looks like
- So what is happening now?

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}$$

$$D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$$

$$(\Lambda + \alpha\mathbb{I})^{-1}\Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\alpha_1} & & & \\ & \frac{\lambda_2}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{\lambda_n}{\lambda_n+\alpha_n} \end{bmatrix}$$

- Each element i of $Q^T w^*$ gets scaled by $\frac{\lambda_i}{\lambda_i+\alpha}$ before it is rotated back by Q

$$\widetilde{w} = Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha \mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}$$

$$D = (\Lambda + \alpha \mathbb{I})^{-1} \Lambda$$

$$(\Lambda + \alpha \mathbb{I})^{-1} \Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\alpha_1} & & & \\ & \frac{\lambda_2}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{\lambda_n}{\lambda_n+\alpha_n} \end{bmatrix}$$

- Each element i of $Q^T w^*$ gets scaled by $\frac{\lambda_i}{\lambda_i+\alpha}$ before it is rotated back by Q
- if $\lambda_i >> \alpha$ then $\frac{\lambda_i}{\lambda_i+\alpha} = 1$

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}$$

$$D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$$

$$(\Lambda + \alpha\mathbb{I})^{-1}\Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\alpha_1} & & & \\ & \frac{\lambda_2}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{\lambda_n}{\lambda_n+\alpha_n} \end{bmatrix}$$

- Each element i of $Q^T w^*$ gets scaled by $\frac{\lambda_i}{\lambda_i+\alpha}$ before it is rotated back by Q
- if $\lambda_i >> \alpha$ then $\frac{\lambda_i}{\lambda_i+\alpha} = 1$
- if $\lambda_i << \alpha$ then $\frac{\lambda_i}{\lambda_i+\alpha} = 0$

$$\widetilde{w} = Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^*$$
$$= QDQ^T w^*$$

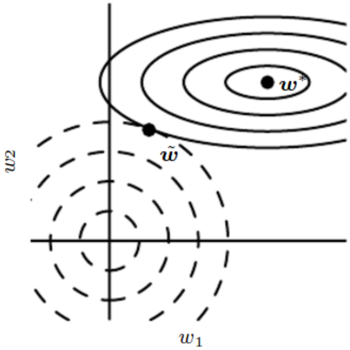
$$(\Lambda + \alpha\mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha_1} & & & \\ & \frac{1}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha_n} \end{bmatrix}$$

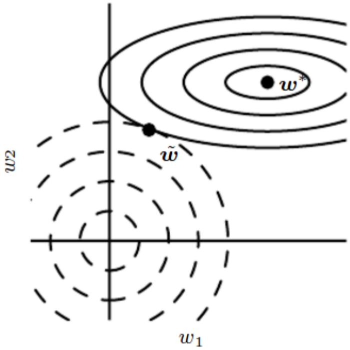
$$D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$$

$$(\Lambda + \alpha\mathbb{I})^{-1}\Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\alpha_1} & & & \\ & \frac{\lambda_2}{\lambda_2+\alpha_2} & & \\ & & \ddots & \\ & & & \frac{\lambda_n}{\lambda_n+\alpha_n} \end{bmatrix}$$

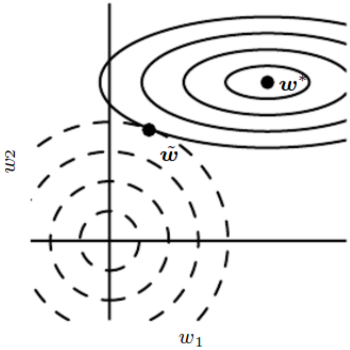
- Each element i of $Q^T w^*$ gets scaled by $\frac{\lambda_i}{\lambda_i+\alpha}$ before it is rotated back by Q
- if $\lambda_i >> \alpha$ then $\frac{\lambda_i}{\lambda_i+\alpha} = 1$
- if $\lambda_i << \alpha$ then $\frac{\lambda_i}{\lambda_i+\alpha} = 0$
- Thus only significant directions (larger eigen values) will be retained.

$$\text{Effective parameters} = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda_i + \alpha} < n$$

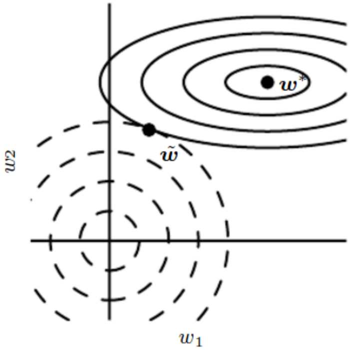

w^*
$\tilde{w}$
w_2
w_1



- The weight vector(w^*) is getting rotated to ($\tilde{w}$)



- The weight vector(w^*) is getting rotated to ($\tilde{w}$)
- All of its elements are shrinking but some are shrinking more than the others



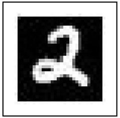
- The weight vector(w^*) is getting rotated to ($\tilde{w}$)
- All of its elements are shrinking but some are shrinking more than the others
- This ensures that only important features are given high weights

Different forms of regularization

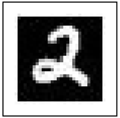
- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

Different forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

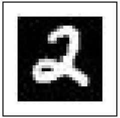


label $= 2$



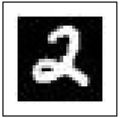
label $= 2$

[given training data]



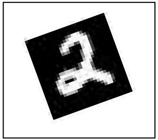
label $= 2$

[given training data]

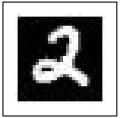


label $= 2$

[given training data]

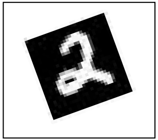


rotated by 20°

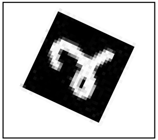


label = 2

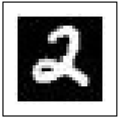
[given training data]



rotated by 20°

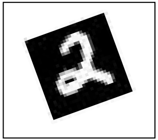


rotated by 65°

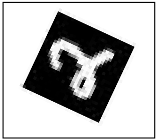


label $= 2$

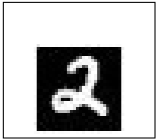
[given training data]



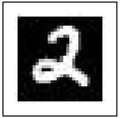
rotated by 20°



rotated by 65°

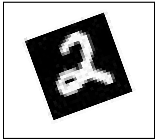


shifted vertically

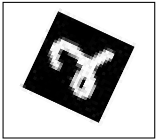


label = 2

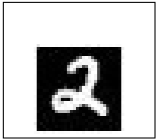
[given training data]



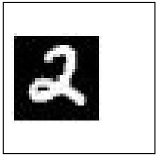
rotated by 20°



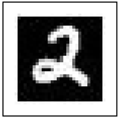
rotated by 65°



shifted vertically

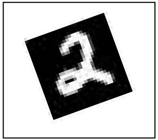


shifted horizontally

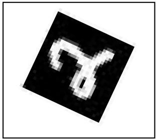


label $= 2$

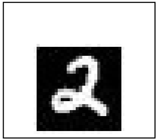
[given training data]



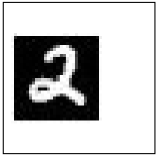
rotated by 20°



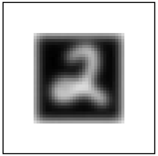
rotated by 65°



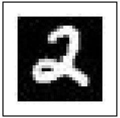
shifted vertically



shifted horizontally

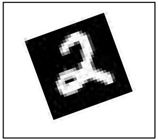


blurred

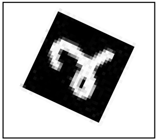


label $= 2$

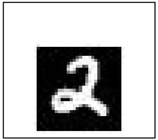
[given training data]



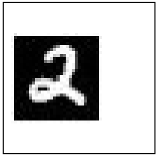
rotated by 20°



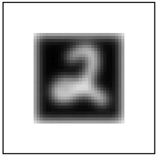
rotated by 65°



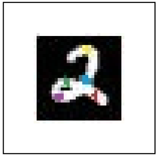
shifted vertically



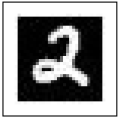
shifted horizontally



blurred

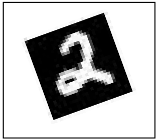


changed some pixels

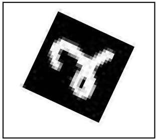


label = 2

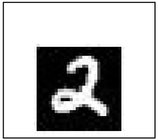
[given training data]



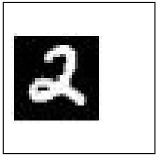
rotated by 20°



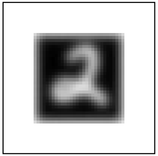
rotated by 65°



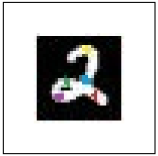
shifted vertically



shifted horizontally

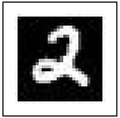


blurred



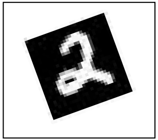
changed some pixels

label = 2

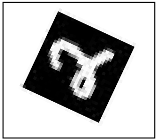


label $= 2$

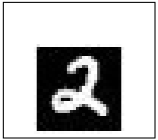
[given training data]



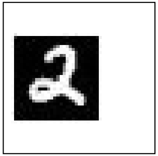
rotated by 20°



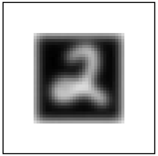
rotated by 65°



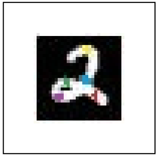
shifted vertically



shifted horizontally



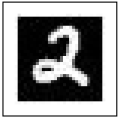
blurred



changed some pixels

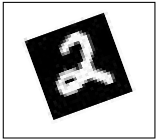
label $= 2$

[augmented data $=$ created using some knowledge of the task]

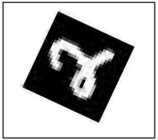


label $= 2$

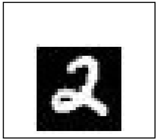
[given training data]
We exploit the fact that certain transformations to the image do not change the label of the image.



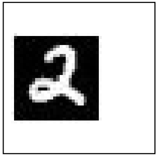
rotated by 20°



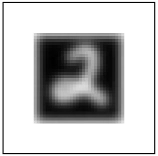
rotated by 65°



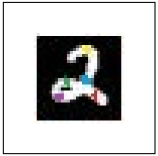
shifted vertically



shifted horizontally



blurred



changed some pixels

label $= 2$

[augmented data $=$ created using some knowledge of the task]

- Typically, More data = better learning

- Typically, More data $=$ better learning
- Works well for image classification / object recognition tasks

- Typically, More data = better learning
- Works well for image classification / object recognition tasks
- Also shown to work well for speech

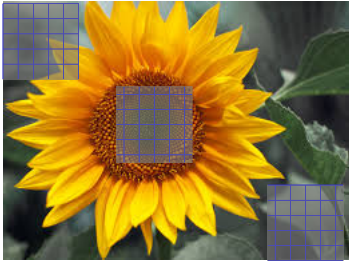
- Typically, More data = better learning
- Works well for image classification / object recognition tasks
- Also shown to work well for speech
- For some tasks it may not be clear how to generate such data

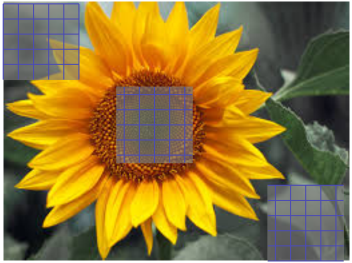
Other forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

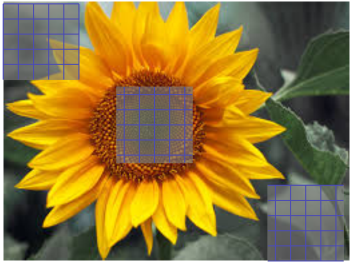
Other forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout



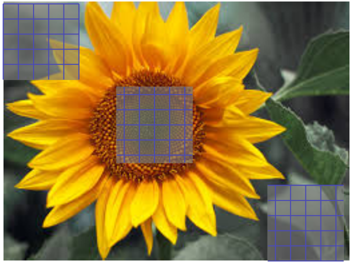


Parameter Sharing



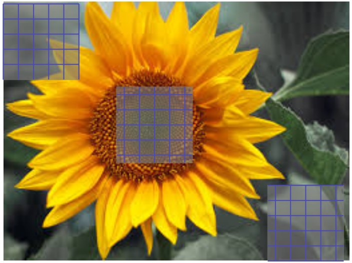
Parameter Sharing

- Used in CNNs



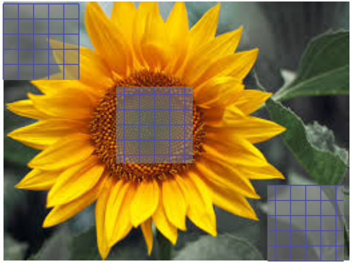
Parameter Sharing

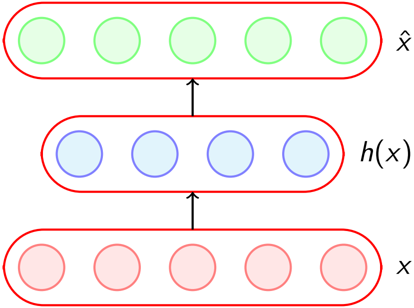
- Used in CNNs
- Same filter applied at different positions of the image



Parameter Sharing

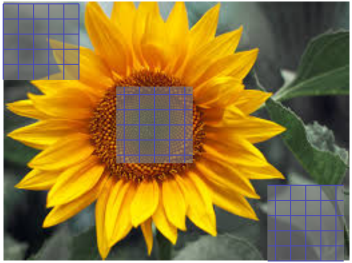
- Used in CNNs
- Same filter applied at different positions of the image
- Or same weight matrix acts on different input neurons

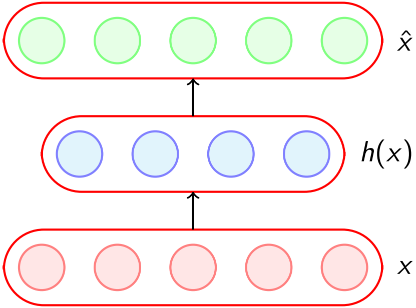




Parameter Sharing

- Used in CNNs
- Same filter applied at different positions of the image
- Or same weight matrix acts on different input neurons

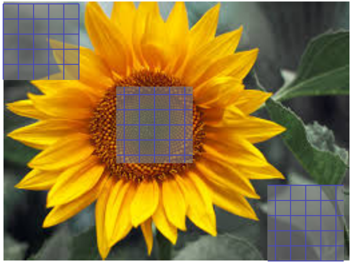


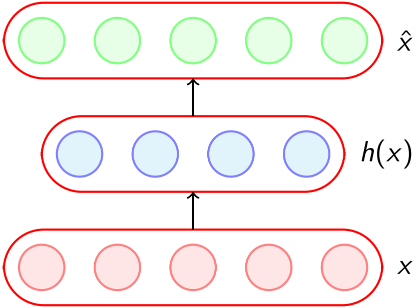


Parameter Sharing

- Used in CNNs
- Same filter applied at different positions of the image
- Or same weight matrix acts on different input neurons

Parameter Tying



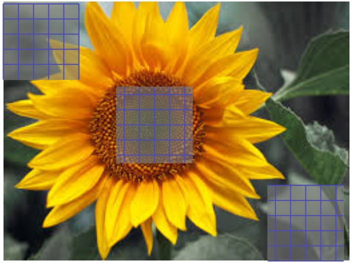


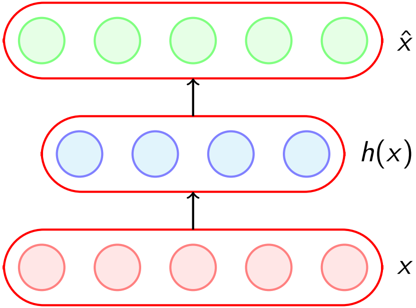
Parameter Sharing

- Used in CNNs
- Same filter applied at different positions of the image
- Or same weight matrix acts on different input neurons

Parameter Tying

- Typically used in autoencoders





Parameter Sharing

- Used in CNNs
- Same filter applied at different positions of the image
- Or same weight matrix acts on different input neurons

Parameter Tying

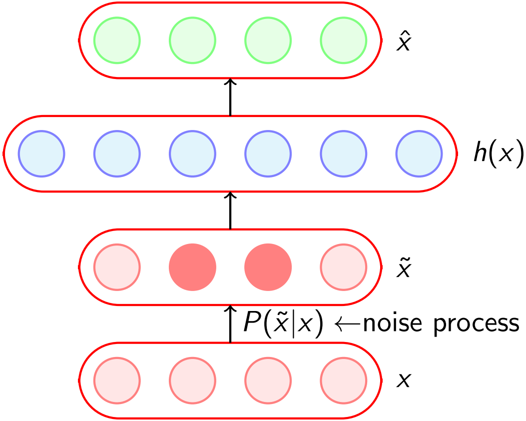
- Typically used in autoencoders
- The encoder and decoder weights are tied.

Other forms of regularization

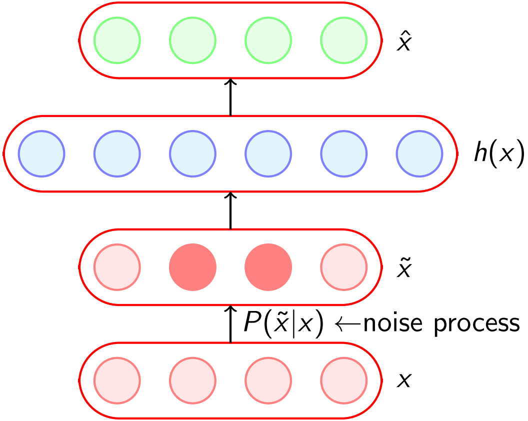
- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

Other forms of regularization

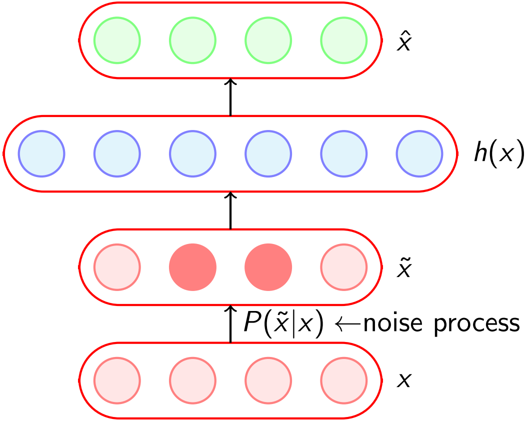
- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout



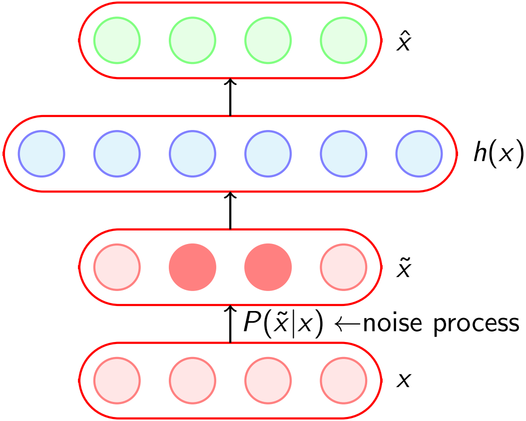
$\hat{x}$
$h(x)$
$\tilde{x}$
$P(\tilde{x}|x)$ ←noise process
x



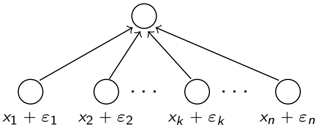
- We saw this in Autoencoder



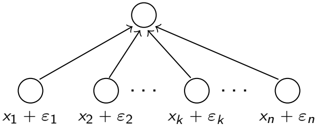
- We saw this in Autoencoder
- We can show that for a simple input output neural network, adding Gaussian noise to the input is equivalent to weight decay ($L2$ normalizaton)



- We saw this in Autoencoder
- We can show that for a simple input output neural network, adding Gaussian noise to the input is equivalent to weight decay ($L2$ normalizaton)
- Can be viewed as data augmentation

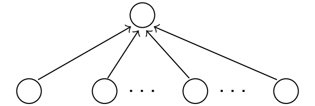


$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$



$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

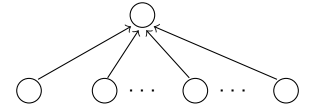
$$\widetilde{x}_i = x_i + \varepsilon_i$$



$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\widetilde{x}_i = x_i + \varepsilon_i$$

$$\widehat{y} = \sum_{i=1}^{n} w_i x_i$$

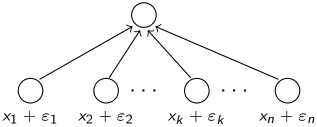


$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\widetilde{x}_i = x_i + \varepsilon_i$$

$$\widehat{y} = \sum_{i=1}^{n} w_i x_i$$

$$\widetilde{y} = \sum_{i=1}^{n} w_i \widetilde{x}_i$$



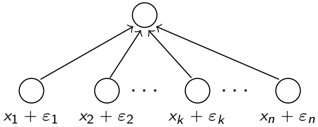
$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\widetilde{x}_i = x_i + \varepsilon_i$$

$$\widehat{y} = \sum_{i=1}^{n} w_i x_i$$

$$\widetilde{y} = \sum_{i=1}^{n} w_i \widetilde{x}_i$$

$$= \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i \varepsilon_i$$

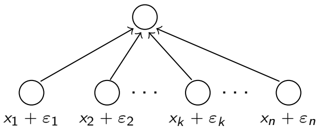


$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\widetilde{x}_i = x_i + \varepsilon_i$$

$$\widehat{y} = \sum_{i=1}^{n} w_i x_i$$

$$\begin{aligned}\widetilde{y} &= \sum_{i=1}^{n} w_i \widetilde{x}_i \\ &= \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i \varepsilon_i \\ &= \widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i\end{aligned}$$



$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\widetilde{x}_i = x_i + \varepsilon_i$$

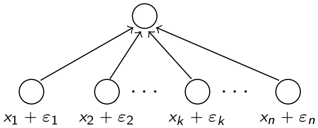
$$\widehat{y} = \sum_{i=1}^{n} w_i x_i$$

$$\widetilde{y} = \sum_{i=1}^{n} w_i \widetilde{x}_i$$

$$= \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i \varepsilon_i$$

$$= \widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i$$

We are interested in $E[(\widetilde{y} - y)^2]$

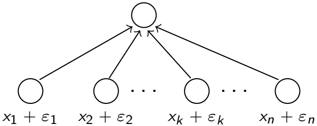


$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\begin{aligned}
\widetilde{x}_i &= x_i + \varepsilon_i \\
\widehat{y} &= \sum_{i=1}^{n} w_i x_i \\
\widetilde{y} &= \sum_{i=1}^{n} w_i \widetilde{x}_i \\
&= \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i \varepsilon_i \\
&= \widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i
\end{aligned}$$

We are interested in $E[(\widetilde{y} - y)^2]$

$$E\left[(\widetilde{y} - y)^2\right] = E\left[\left(\widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i - y\right)^2\right]$$



$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\widetilde{x}_i = x_i + \varepsilon_i$$

$$\widehat{y} = \sum_{i=1}^{n} w_i x_i$$

$$\widetilde{y} = \sum_{i=1}^{n} w_i \widetilde{x}_i$$

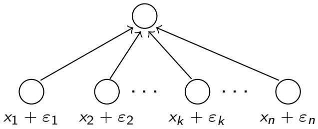
$$= \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i \varepsilon_i$$

$$= \widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i$$

We are interested in $E[(\widetilde{y} - y)^2]$

$$E\left[(\widetilde{y} - y)^2\right] = E\left[\left(\widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i - y\right)^2\right]$$

$$= E\left[\left(\left(\widehat{y} - y\right) + \left(\sum_{i=1}^{n} w_i \varepsilon_i\right)\right)^2\right]$$



$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\widetilde{x}_i = x_i + \varepsilon_i$$

$$\widehat{y} = \sum_{i=1}^{n} w_i x_i$$

$$\widetilde{y} = \sum_{i=1}^{n} w_i \widetilde{x}_i$$

$$= \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i \varepsilon_i$$

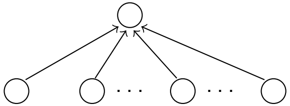
$$= \widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i$$

We are interested in $E[(\widetilde{y} - y)^2]$

$$E\left[(\widetilde{y} - y)^2\right] = E\left[\left(\widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i - y\right)^2\right]$$

$$= E\left[\left(\left(\widehat{y} - y\right) + \left(\sum_{i=1}^{n} w_i \varepsilon_i\right)\right)^2\right]$$

$$= E\left[(\widehat{y} - y)^2\right] + E\left[2(\widehat{y} - y)\sum_{i=1}^{n} w_i \varepsilon_i\right] + E\left[\left(\sum_{i=1}^{n} w_i \varepsilon_i\right)^2\right]$$



$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\widetilde{x}_i = x_i + \varepsilon_i$$

$$\widehat{y} = \sum_{i=1}^{n} w_i x_i$$

$$\widetilde{y} = \sum_{i=1}^{n} w_i \widetilde{x}_i$$

$$= \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i \varepsilon_i$$

$$= \widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i$$

We are interested in $E[(\widetilde{y} - y)^2]$

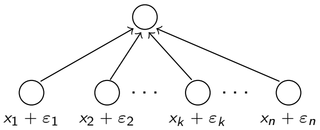
$$E\left[(\widetilde{y} - y)^2\right] = E\left[\left(\widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i - y\right)^2\right]$$

$$= E\left[\left(\left(\widehat{y} - y\right) + \left(\sum_{i=1}^{n} w_i \varepsilon_i\right)\right)^2\right]$$

$$= E\left[(\widehat{y} - y)^2\right] + E\left[2(\widehat{y} - y)\sum_{i=1}^{n} w_i \varepsilon_i\right] + E\left[\left(\sum_{i=1}^{n} w_i \varepsilon_i\right)^2\right]$$

$$= E\left[(\widehat{y} - y)^2\right] + 0 + E\left[\sum_{i=1}^{n} w_i^2 \varepsilon_i^2\right]$$

($\because$ ε_i is independent of ε_j and ε_i is independent of ($\widehat{y}$-y))



$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\widetilde{x}_i = x_i + \varepsilon_i$$

$$\widehat{y} = \sum_{i=1}^{n} w_i x_i$$

$$\widetilde{y} = \sum_{i=1}^{n} w_i \widetilde{x}_i$$

$$= \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i \varepsilon_i$$

$$= \widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i$$

We are interested in $E[(\widetilde{y} - y)^2]$

$$E\left[(\widetilde{y} - y)^2\right] = E\left[\left(\widehat{y} + \sum_{i=1}^{n} w_i \varepsilon_i - y\right)^2\right]$$

$$= E\left[\left(\left(\widehat{y} - y\right) + \left(\sum_{i=1}^{n} w_i \varepsilon_i\right)\right)^2\right]$$

$$= E\left[(\widehat{y} - y)^2\right] + E\left[2(\widehat{y} - y)\sum_{i=1}^{n} w_i \varepsilon_i\right] + E\left[\left(\sum_{i=1}^{n} w_i \varepsilon_i\right)^2\right]$$

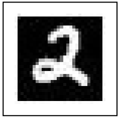
$$= E\left[(\widehat{y} - y)^2\right] + 0 + E\left[\sum_{i=1}^{n} w_i^2 \varepsilon_i^2\right]$$

($\because$ ε_i is independent of ε_j and ε_i is independent of ($\widehat{y}$-y))

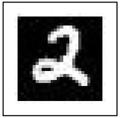
$$= (E\left[(\widehat{y} - y)^2\right] + \sigma^2 \sum_{i=1}^{n} w_i^2 \quad \text{(same as L2 norm penalty)}$$

Other forms of regularization

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

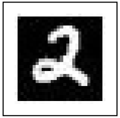


0	0	1	0	0	0	0	0	0	0	Hard targets



0	0	1	0	0	0	0	0	0	0	Hard targets

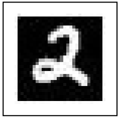
$$\text{minimize} : \sum_{i=0}^{9} p_i \log q_i$$



0	0	1	0	0	0	0	0	0	0	Hard targets

$$\text{minimize} : \sum_{i=0}^{9} p_i \log q_i$$

$$\text{true distribution} : p = \{0, 0, 1, 0, 0, 0, 0, 0, 0, 0\}$$

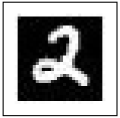


0	0	1	0	0	0	0	0	0	0	Hard targets

$$\text{minimize} : \sum_{i=0}^{9} p_i \log q_i$$

$$\text{true distribution} : p = \{0, 0, 1, 0, 0, 0, 0, 0, 0, 0\}$$

$$\text{estimated distribution} : q$$



0	0	1	0	0	0	0	0	0	0	Hard targets

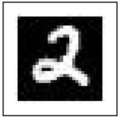
$$\text{minimize}: \sum_{i=0}^{9} p_i \log q_i$$

$$\text{true distribution}: p = \{0, 0, 1, 0, 0, 0, 0, 0, 0, 0\}$$

$$\text{estimated distribution}: q$$

Intuition

- Do not trust the true labels, they may be noisy



0	0	1	0	0	0	0	0	0	0	Hard targets

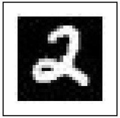
$$\text{minimize} : \sum_{i=0}^{9} p_i \log q_i$$

$$\text{true distribution} : p = \{0, 0, 1, 0, 0, 0, 0, 0, 0, 0\}$$

$$\text{estimated distribution} : q$$

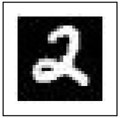
Intuition

- Do not trust the true labels, they may be noisy
- Instead, use soft targets



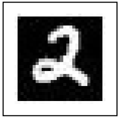
$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$1-\varepsilon$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$

Soft targets



$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$1-\varepsilon$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	Soft targets

$\varepsilon =$ small positive constant

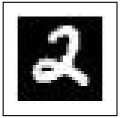


$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$1-\varepsilon$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$

Soft targets

ε = small positive constant

$$\text{minimize} : \sum_{i=0}^{9} p_i \log q_i$$

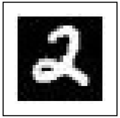


$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$1-\varepsilon$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	Soft targets

$$\varepsilon = \text{small positive constant}$$

$$\text{minimize} : \sum_{i=0}^{9} p_i \log q_i$$

$$\text{true distribution} + \text{noise} : p = \left\{\frac{\varepsilon}{9}, \frac{\varepsilon}{9}, 1-\varepsilon, \frac{\varepsilon}{9}, \ldots\right\}$$



$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$1-\varepsilon$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	$\frac{\varepsilon}{9}$	Soft targets

$\varepsilon =$ small positive constant

$$\text{minimize} : \sum_{i=0}^{9} p_i \log q_i$$

$$\text{true distribution} + \text{noise} : p = \left\{ \frac{\varepsilon}{9}, \frac{\varepsilon}{9}, 1-\varepsilon, \frac{\varepsilon}{9}, \dots \right\}$$

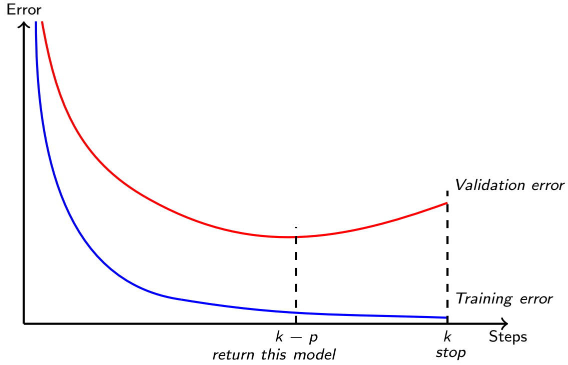
$$\text{estimated distribution} : q$$

Other forms of regularization

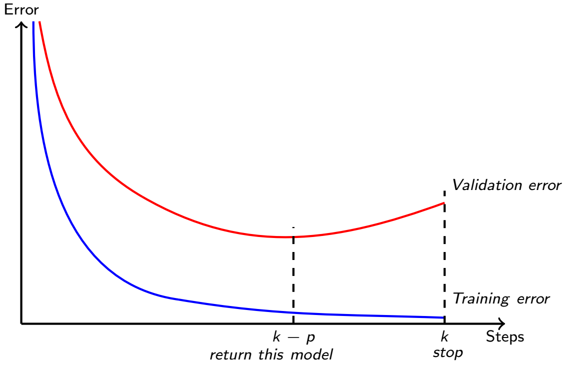
- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

Other forms of regularization

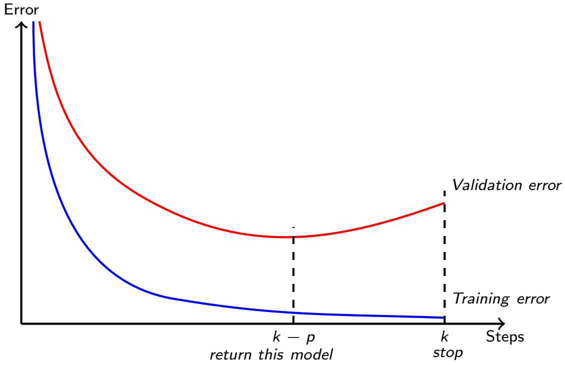
- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout



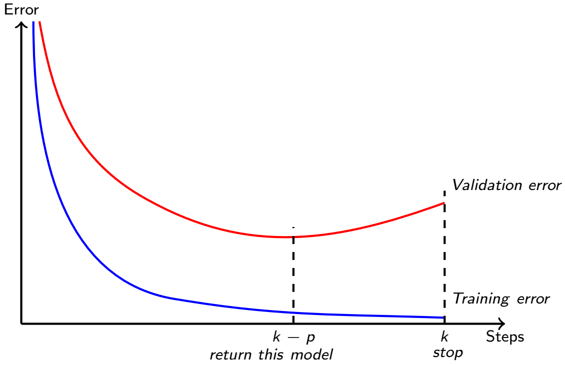
- Track the validation error



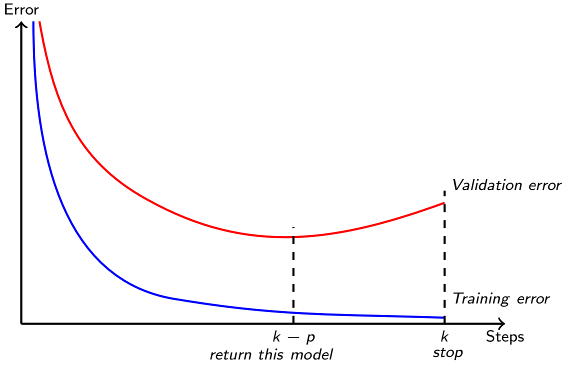
- Track the validation error
- Have a patience parameter p



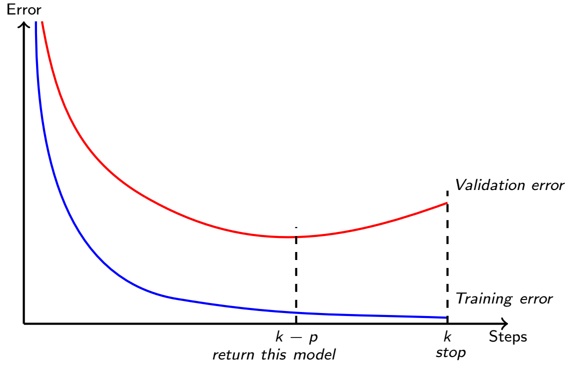
- Track the validation error
- Have a patience parameter p
- If you are at step k and there was no improvement in validation error in the previous p steps then stop training and return the model stored at step $k - p$



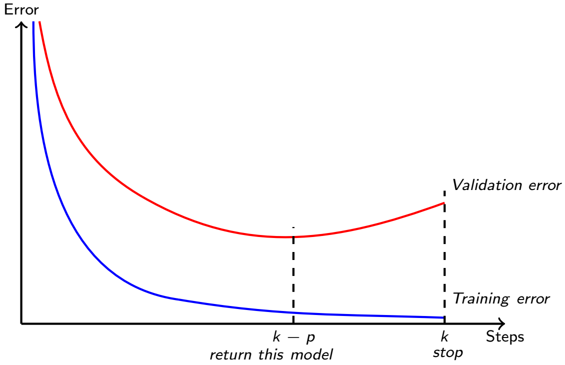
- Track the validation error
- Have a patience parameter p
- If you are at step k and there was no improvement in validation error in the previous p steps then stop training and return the model stored at step $k - p$
- Basically, stop the training early before it drives the training error to 0 and blows up the validation error



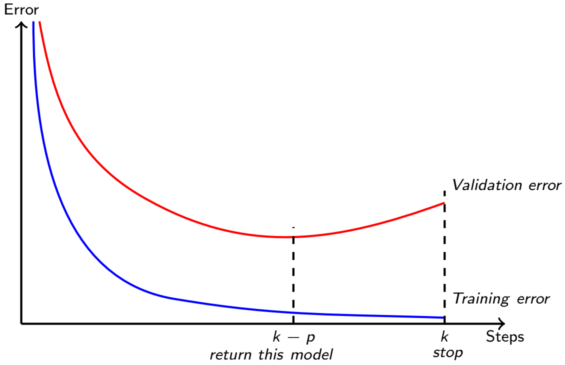
- Very effective and the mostly widely used form of regularization



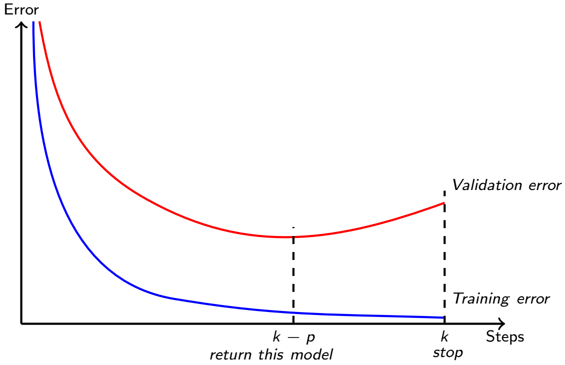
- Very effective and the mostly widely used form of regularization
- Can be used even with other regularizers(such as L_2)



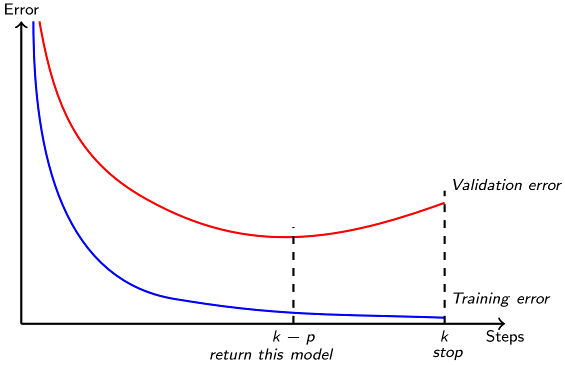
- Very effective and the mostly widely used form of regularization
- Can be used even with other regularizers(such as L_2)
- How does it act as a regularizer ?



- Very effective and the mostly widely used form of regularization
- Can be used even with other regularizers(such as L_2)
- How does it act as a regularizer ?
- We will first see an intuitive explanation and then a mathematical analysis

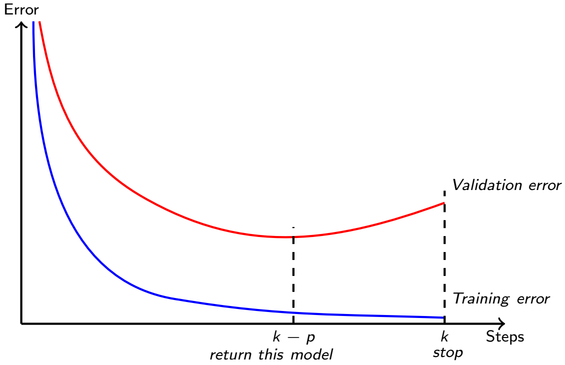


- Recall that the update rule in SGD is-



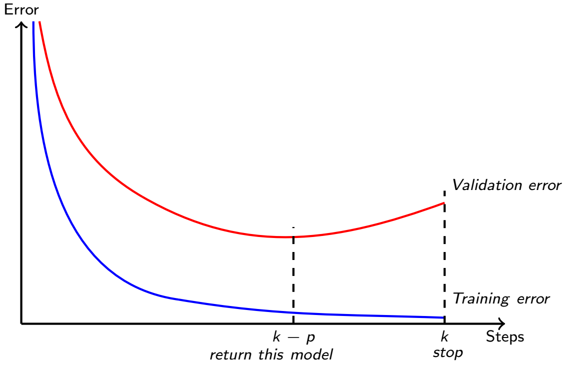
- Recall that the update rule in SGD is-

$$\omega_{t+1} = \omega_t + \eta \nabla \omega_t$$



- Recall that the update rule in SGD is-

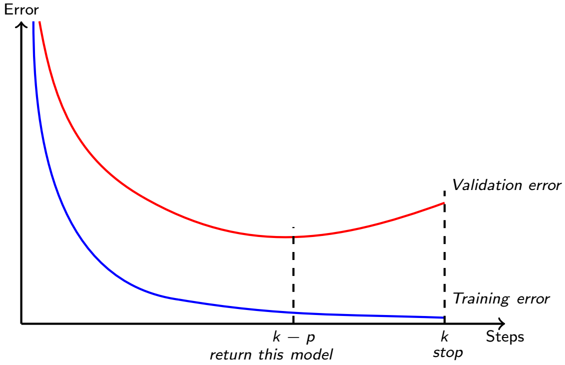
$$\omega_{t+1} = \omega_t + \eta \nabla \omega_t$$
$$= \omega_0 + \eta \sum_{i=1}^{t} \nabla \omega_i$$



- Recall that the update rule in SGD is-

$$\begin{aligned}\omega_{t+1} &= \omega_t + \eta \nabla \omega_t \\ &= \omega_0 + \eta \sum_{i=1}^{t} \nabla \omega_i\end{aligned}$$

- Let τ be the maximum value of $\nabla \omega_i$ then

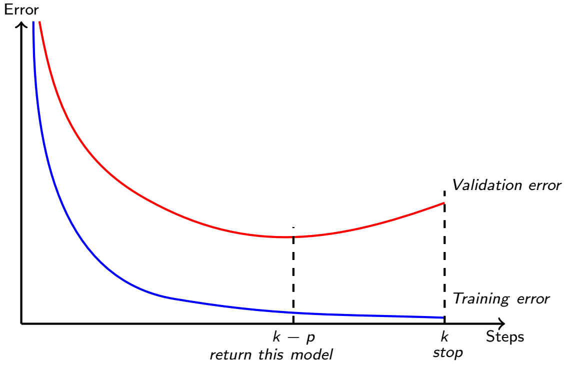


- Recall that the update rule in SGD is-

$$\omega_{t+1} = \omega_t + \eta \nabla \omega_t$$
$$= \omega_0 + \eta \sum_{i=1}^{t} \nabla \omega_i$$

- Let τ be the maximum value of $\nabla \omega_i$ then

$$\omega_t \leq \omega_0 + \eta t \tau$$



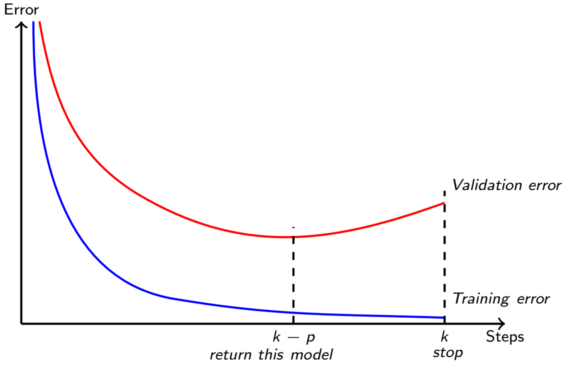
- Recall that the update rule in SGD is-

$$\omega_{t+1} = \omega_t + \eta \nabla \omega_t$$
$$= \omega_0 + \eta \sum_{i=1}^{t} \nabla \omega_i$$

- Let τ be the maximum value of $\nabla \omega_i$ then

$$\omega_t \leq \omega_0 + \eta t \tau$$

- Thus, t controls how far ω_t can go from the initial ω_0



- Recall that the update rule in SGD is-

$$\omega_{t+1} = \omega_t + \eta \nabla \omega_t$$
$$= \omega_0 + \eta \sum_{i=1}^{t} \nabla \omega_i$$

- Let τ be the maximum value of $\nabla \omega_i$ then

$$\omega_t \leq \omega_0 + \eta t \tau$$

- Thus, t controls how far ω_t can go from the initial ω_0
- In other words it controls the space of exploration

We will now see a mathematical analysis of this

- Recall that the Taylor series approximation for $J(\omega)$ is

- Recall that the Taylor series approximation for $J(\omega)$ is

$$J(\omega) = J(\omega^*) + (\omega - \omega^*)^T \nabla J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*)$$

- Recall that the Taylor series approximation for $J(\omega)$ is

$$
\begin{aligned}
J(\omega) &= J(\omega^*) + (\omega - \omega^*)^T \nabla J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \\
&= J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \qquad [\ \omega^* \textit{ is optimal so } \nabla J(\omega^*) \textit{ is } 0\ ]
\end{aligned}
$$

- Recall that the Taylor series approximation for $J(\omega)$ is

$$J(\omega) = J(\omega^*) + (\omega - \omega^*)^T \nabla J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*)$$

$$= J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \qquad [\ \omega^* \textit{ is optimal so } \nabla J(\omega^*) \textit{ is } 0\ ]$$

$$\nabla(J(\omega)) = H(\omega - \omega^*)$$

- Recall that the Taylor series approximation for $J(\omega)$ is

$$J(\omega) = J(\omega^*) + (\omega - \omega^*)^T \nabla J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*)$$

$$= J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \qquad [\ \omega^* \textit{ is optimal so } \nabla J(\omega^*) \textit{ is } 0\ ]$$

$$\nabla(J(\omega)) = H(\omega - \omega^*)$$

Now the SGD update rule is:

- Recall that the Taylor series approximation for $J(\omega)$ is

$$\begin{aligned} J(\omega) &= J(\omega^*) + (\omega - \omega^*)^T \nabla J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \\ &= J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \qquad [\ \omega^* \textit{ is optimal so } \nabla J(\omega^*) \textit{ is } 0\ ] \\ \nabla(J(\omega)) &= H(\omega - \omega^*) \end{aligned}$$

Now the SGD update rule is:

$$\omega_t = \omega_{t-1} + \eta \nabla J(\omega_{t-1})$$

- Recall that the Taylor series approximation for $J(\omega)$ is

$$\begin{aligned}
J(\omega) &= J(\omega^*) + (\omega - \omega^*)^T \nabla J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \\
&= J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \qquad [\ \omega^* \ \textit{is optimal so}\ \nabla J(\omega^*)\ \textit{is}\ 0\ ] \\
\nabla(J(\omega)) &= H(\omega - \omega^*)
\end{aligned}$$

Now the SGD update rule is:

$$\begin{aligned}
\omega_t &= \omega_{t-1} + \eta \nabla J(\omega_{t-1}) \\
&= \omega_{t-1} + \eta H(\omega_{t-1} - \omega^*)
\end{aligned}$$

- Recall that the Taylor series approximation for $J(\omega)$ is

$$
\begin{aligned}
J(\omega) &= J(\omega^*) + (\omega - \omega^*)^T \nabla J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \\
&= J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \qquad [\ \omega^* \textit{ is optimal so } \nabla J(\omega^*) \textit{ is } 0\ ] \\
\nabla(J(\omega)) &= H(\omega - \omega^*)
\end{aligned}
$$

Now the SGD update rule is:

$$
\begin{aligned}
\omega_t &= \omega_{t-1} + \eta \nabla J(\omega_{t-1}) \\
&= \omega_{t-1} + \eta H(\omega_{t-1} - \omega^*) \\
&= (I + \eta H)\omega_{t-1} - \eta H \omega^*
\end{aligned}
$$

$$\omega_t = (I + \eta H)\omega_{t-1} - \eta H \omega^*$$

$$\omega_t = (I + \eta H)\omega_{t-1} - \eta H\omega^*$$

- Using EVD of H as $H = Q\Lambda Q^T$, we get:

$$\omega_t = (I + \eta Q\Lambda Q^T)\omega_{t-1} - \eta Q\Lambda Q^T\omega^*$$

$$\omega_t = (I + \eta H)\omega_{t-1} - \eta H \omega^*$$

- Using EVD of H as $H = Q\Lambda Q^T$, we get:

$$\omega_t = (I + \eta Q\Lambda Q^T)\omega_{t-1} - \eta Q\Lambda Q^T \omega^*$$

- If we start with $\omega_0 = 0$ then we can show that

$$\omega_t = Q[I - (I - \varepsilon\Lambda)^t]Q^T \omega^*$$

$$\omega_t = (I + \eta H)\omega_{t-1} - \eta H\omega^*$$

- Using EVD of H as $H = Q\Lambda Q^T$, we get:

$$\omega_t = (I + \eta Q\Lambda Q^T)\omega_{t-1} - \eta Q\Lambda Q^T\omega^*$$

- If we start with $\omega_0 = 0$ then we can show that

$$\omega_t = Q[I - (I - \varepsilon\Lambda)^t]Q^T\omega^*$$

- Compare this with the expression we had for optimum $\tilde{\omega}$ with L_2 regularization

$$\tilde{\omega} = Q[I - (\Lambda + \alpha I)^{-1}\alpha]Q^T\omega^*$$

$$\omega_t = (I + \eta H)\omega_{t-1} - \eta H \omega^*$$

- Using EVD of H as $H = Q\Lambda Q^T$, we get:

$$\omega_t = (I + \eta Q\Lambda Q^T)\omega_{t-1} - \eta Q\Lambda Q^T \omega^*$$

- If we start with $\omega_0 = 0$ then we can show that

$$\omega_t = Q[I - (I - \varepsilon\Lambda)^t]Q^T\omega^*$$

- Compare this with the expression we had for optimum $\tilde{\omega}$ with L_2 regularization

$$\tilde{\omega} = Q[I - (\Lambda + \alpha I)^{-1}\alpha]Q^T\omega^*$$

- We observe that $\omega_t = \tilde{\omega}$, if we choose ε,t and α such that

$$(I - \varepsilon\Lambda)^t = (\Lambda + \alpha I)^{-1}\alpha$$

Things to be remember

- Early stopping only allows t updates to the parameters.

Things to be remember

- Early stopping only allows t updates to the parameters.
- If a parameter ω corresponds to a dimension which is important for the loss $\mathscr{L}(\theta)$ then $\frac{\partial \mathscr{L}(\theta)}{\partial \omega}$ will be large

Things to be remember

- Early stopping only allows t updates to the parameters.
- If a parameter ω corresponds to a dimension which is important for the loss $\mathscr{L}(\theta)$ then $\frac{\partial \mathscr{L}(\theta)}{\partial \omega}$ will be large

Things to be remember

- Early stopping only allows t updates to the parameters.
- If a parameter ω corresponds to a dimension which is important for the loss $\mathscr{L}(\theta)$ then $\frac{\partial \mathscr{L}(\theta)}{\partial \omega}$ will be large
- However if a parameter is not important ($\frac{\partial \mathscr{L}(\theta)}{\partial \omega}$ is small) then its updates will be small and the parameter will not be able to grow large in 't' steps

Things to be remember

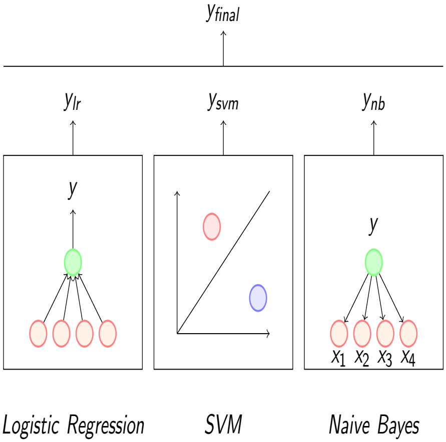
- Early stopping only allows t updates to the parameters.
- If a parameter ω corresponds to a dimension which is important for the loss $\mathscr{L}(\theta)$ then $\frac{\partial \mathscr{L}(\theta)}{\partial \omega}$ will be large
- However if a parameter is not important ($\frac{\partial \mathscr{L}(\theta)}{\partial \omega}$ is small) then its updates will be small and the parameter will not be able to grow large in 't' steps
- Early stopping will thus effectively shrink the parameters corresponding to less important directions (same as weight decay).

Other forms of regularization

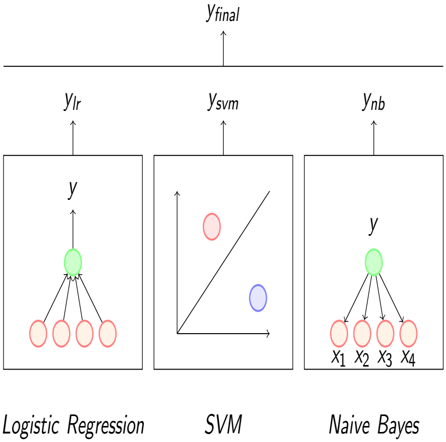
- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

Other forms of regularization

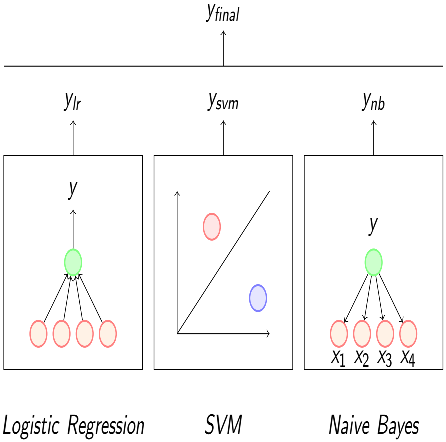
- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout



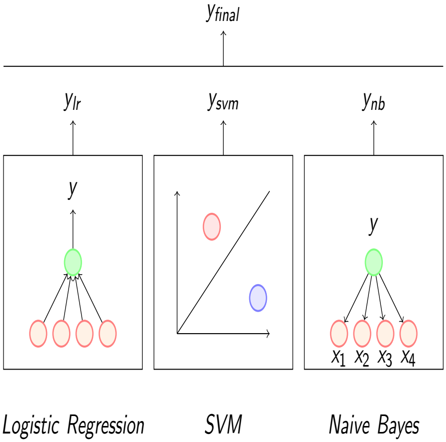
- Combine the output of different models to reduce generalization error



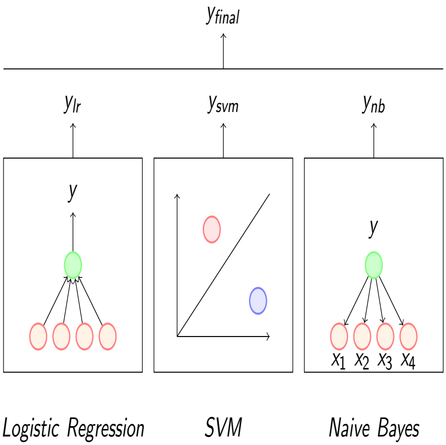
- Combine the output of different models to reduce generalization error
- The models can correspond to different classifiers



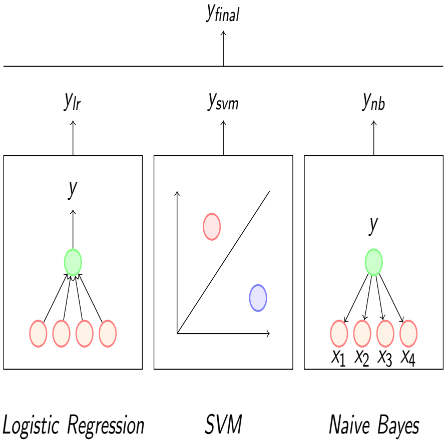
- Combine the output of different models to reduce generalization error
- The models can correspond to different classifiers
- It could be different instances of the same classifier trained with:



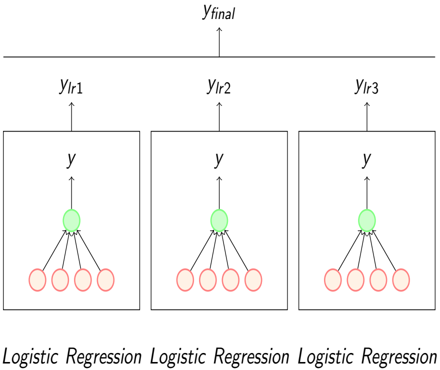
- Combine the output of different models to reduce generalization error
- The models can correspond to different classifiers
- It could be different instances of the same classifier trained with:
  - different hyperparameters

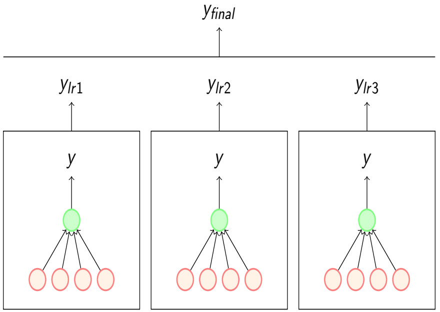


- Combine the output of different models to reduce generalization error
- The models can correspond to different classifiers
- It could be different instances of the same classifier trained with:
  - different hyperparameters
  - different features

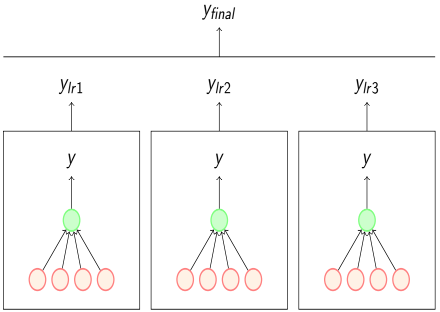


- Combine the output of different models to reduce generalization error
- The models can correspond to different classifiers
- It could be different instances of the same classifier trained with:
  - different hyperparameters
  - different features
  - different samples of the training data


y_{final}
y_{lr1}
y_{lr2}
y_{lr3}
y
y
y
Logistic Regression
Logistic Regression
Logistic Regression

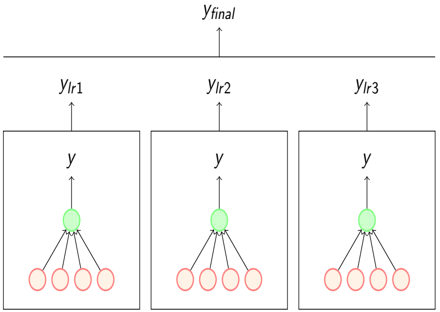


Logistic Regression *Logistic Regression* *Logistic Regression*



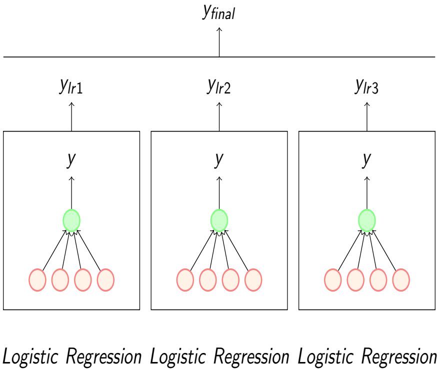
Logistic Regression *Logistic Regression* *Logistic Regression*

- Bagging: form an ensemble using different instances of the same classifier

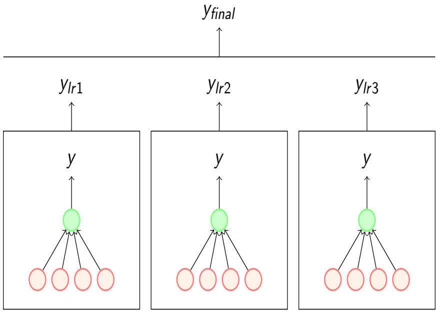


Logistic Regression Logistic Regression Logistic Regression

- Bagging: form an ensemble using different instances of the same classifier
- From a given dataset, construct multiple training sets by sampling with replacement ($T_1, T_2, ..., T_k$)



- Bagging: form an ensemble using different instances of the same classifier
- From a given dataset, construct multiple training sets by sampling with replacement ($T_1, T_2, ..., T_k$)
- Train i^{th} instance of the classifier using training set T_i



Logistic Regression *Logistic Regression* *Logistic Regression*

Each model trained with a different sample of the data (sampling with replacement)

- Bagging: form an ensemble using different instances of the same classifier
- From a given dataset, construct multiple training sets by sampling with replacement ($T_1, T_2, ..., T_k$)
- Train i^{th} instance of the classifier using training set T_i

- When would bagging work?

- When would bagging work?
- Consider a set of k LR models

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution
- $Variance = E[\varepsilon_i^2] = V$

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution
- $Variance = E[\varepsilon_i^2] = V$
- $Covariance = E[\varepsilon_i \varepsilon_j] = C$

- The error made by the average prediction of all the models is $\frac{1}{k}\sum_i \varepsilon_i$.

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution
- $Variance = E[\varepsilon_i^2] = V$
- $Covariance = E[\varepsilon_i \varepsilon_j] = C$

- The error made by the average prediction of all the models is $\frac{1}{k}\sum_i \varepsilon_i$.
- The expected squared error is given by:

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution
- $Variance = E[\varepsilon_i^2] = V$
- $Covariance = E[\varepsilon_i \varepsilon_j] = C$

- The error made by the average prediction of all the models is $\frac{1}{k}\sum_i \varepsilon_i$.
- The expected squared error is given by:

$$mse = E[(\frac{1}{k}\sum_i \varepsilon_i)^2]$$

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution
- $Variance = E[\varepsilon_i^2] = V$
- $Covariance = E[\varepsilon_i \varepsilon_j] = C$

- The error made by the average prediction of all the models is $\frac{1}{k}\sum_i \varepsilon_i$.
- The expected squared error is given by:

$$
\begin{aligned}
mse =& E[(\frac{1}{k}\sum_i \varepsilon_i)^2] \\
=& \frac{1}{k^2} E[\sum_i \sum_{i=j} \varepsilon_i \varepsilon_j + \sum_i \sum_{i \neq j} \varepsilon_i \varepsilon_j]
\end{aligned}
$$

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution
- $Variance = E[\varepsilon_i^2] = V$
- $Covariance = E[\varepsilon_i \varepsilon_j] = C$

- The error made by the average prediction of all the models is $\frac{1}{k}\sum_i \varepsilon_i$.
- The expected squared error is given by:

$$mse = E[(\frac{1}{k}\sum_i \varepsilon_i)^2]$$

$$= \frac{1}{k^2} E[\sum_i \sum_{i=j} \varepsilon_i \varepsilon_j + \sum_i \sum_{i \neq j} \varepsilon_i \varepsilon_j]$$

$$= \frac{1}{k^2} E[\sum_i \varepsilon_i^2 + \sum_i \sum_{i \neq j} \varepsilon_i \varepsilon_j]$$

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution
- $Variance = E[\varepsilon_i^2] = V$
- $Covariance = E[\varepsilon_i \varepsilon_j] = C$

- The error made by the average prediction of all the models is $\frac{1}{k}\sum_i \varepsilon_i$.
- The expected squared error is given by:

$$
\begin{aligned}
mse &= E[(\frac{1}{k}\sum_i \varepsilon_i)^2] \\
&= \frac{1}{k^2}E[\sum_i \sum_{i=j} \varepsilon_i \varepsilon_j + \sum_i \sum_{i \neq j} \varepsilon_i \varepsilon_j] \\
&= \frac{1}{k^2}E[\sum_i \varepsilon_i^2 + \sum_i \sum_{i \neq j} \varepsilon_i \varepsilon_j] \\
&= \frac{1}{k^2}(\sum_i^k E[\varepsilon_i^2] + \sum_i^k \sum_{i \neq j} E[\varepsilon_i \varepsilon_j])
\end{aligned}
$$

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution
- $Variance = E[\varepsilon_i^2] = V$
- $Covariance = E[\varepsilon_i \varepsilon_j] = C$

- The error made by the average prediction of all the models is $\frac{1}{k}\sum_i \varepsilon_i$.
- The expected squared error is given by:

$$
\begin{aligned}
mse &= E[(\frac{1}{k}\sum_i \varepsilon_i)^2] \\
&= \frac{1}{k^2}E[\sum_i \sum_{i=j} \varepsilon_i \varepsilon_j + \sum_i \sum_{i \neq j} \varepsilon_i \varepsilon_j] \\
&= \frac{1}{k^2}E[\sum_i \varepsilon_i^2 + \sum_i \sum_{i \neq j} \varepsilon_i \varepsilon_j] \\
&= \frac{1}{k^2}(\sum_i^k E[\varepsilon_i^2] + \sum_i^k \sum_{i \neq j} E[\varepsilon_i \varepsilon_j]) \\
&= \frac{1}{k^2}(kV + k(k-1)C)
\end{aligned}
$$

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution
- $Variance = E[\varepsilon_i^2] = V$
- $Covariance = E[\varepsilon_i \varepsilon_j] = C$

- The error made by the average prediction of all the models is $\frac{1}{k}\sum_i \varepsilon_i$.
- The expected squared error is given by:

$$
\begin{aligned}
mse =& E[(\frac{1}{k}\sum_i \varepsilon_i)^2] \\
=& \frac{1}{k^2}E[\sum_i \sum_{i=j} \varepsilon_i \varepsilon_j + \sum_i \sum_{i \neq j} \varepsilon_i \varepsilon_j] \\
=& \frac{1}{k^2}E[\sum_i \varepsilon_i^2 + \sum_i \sum_{i \neq j} \varepsilon_i \varepsilon_j] \\
=& \frac{1}{k^2}(\sum_i^k E[\varepsilon_i^2] + \sum_i^k \sum_{i \neq j} E[\varepsilon_i \varepsilon_j]) \\
=& \frac{1}{k^2}(kV + k(k-1)C) \\
=& \frac{1}{k}V + \frac{k-1}{k}C
\end{aligned}
$$

- When would bagging work?
- Consider a set of k LR models
- Suppose that each model makes an error ε_i on a test example
- Let ε_i be drawn from a zero mean multivariate normal distribution
- $Variance = E[\varepsilon_i^2] = V$
- $Covariance = E[\varepsilon_i \varepsilon_j] = C$

$$mse = \frac{1}{k}V + \frac{k-1}{k}C$$

$$mse = \frac{1}{k}V + \frac{k-1}{k}C$$

- When would bagging work ?

$$mse = \frac{1}{k}V + \frac{k-1}{k}C$$

- When would bagging work ?
- If the errors of the model are perfectly correlated then $V = C$ and $mse = V$ [bagging does not help: the mse of the ensemble is as bad as the individual models]

$$mse = \frac{1}{k}V + \frac{k-1}{k}C$$

- When would bagging work ?
- If the errors of the model are perfectly correlated then $V = C$ and $mse = V$ [bagging does not help: the mse of the ensemble is as bad as the individual models]
- If the errors of the model are independent or uncorrelated then $C = 0$ and the mse of the ensemble reduces to $\frac{1}{k}V$

$$mse = \frac{1}{k}V + \frac{k-1}{k}C$$

- When would bagging work ?
- If the errors of the model are perfectly correlated then $V = C$ and $mse = V$ [bagging does not help: the mse of the ensemble is as bad as the individual models]
- If the errors of the model are independent or uncorrelated then $C = 0$ and the mse of the ensemble reduces to $\frac{1}{k}V$
- On average, the ensemble will perform at least as well as its individual members

Other forms of regularization

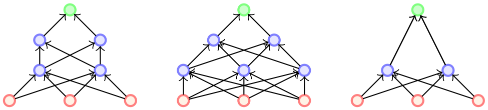
- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

Other forms of regularization

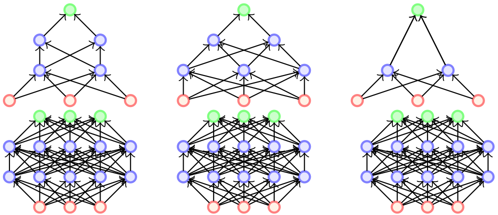
- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

- Typically model averaging(bagging ensemble) always helps

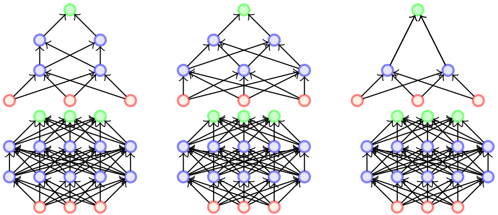
- Typically model averaging(bagging ensemble) always helps
- Training several large neural networks for making an ensemble is prohibitively expensive



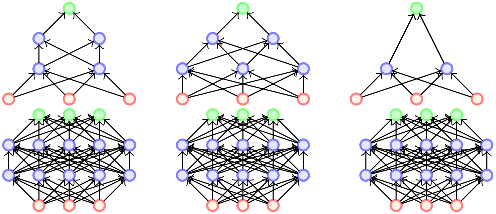
- Typically model averaging(bagging ensemble) always helps
- Training several large neural networks for making an ensemble is prohibitively expensive
- Option 1: Train several neural networks having different architectures(obviously expensive)



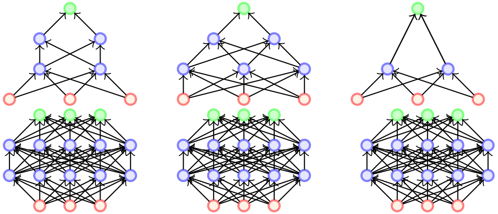
- Typically model averaging(bagging ensemble) always helps
- Training several large neural networks for making an ensemble is prohibitively expensive
- Option 1: Train several neural networks having different architectures(obviously expensive)
- Option 2: Train multiple instances of the same network using different training samples(again expensive)



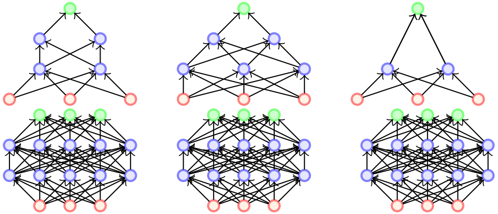
- Typically model averaging(bagging ensemble) always helps
- Training several large neural networks for making an ensemble is prohibitively expensive
- Option 1: Train several neural networks having different architectures(obviously expensive)
- Option 2: Train multiple instances of the same network using different training samples(again expensive)
- Even if we manage to train with option 1 or option 2, combining several models at test time is infeasible in real time applications



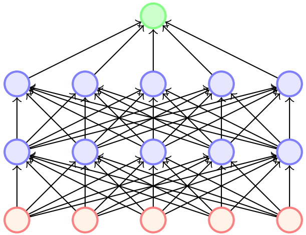
- Dropout is a technique which addresses both these issues.



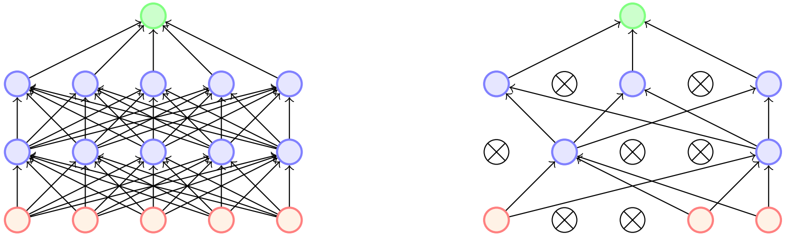
- Dropout is a technique which addresses both these issues.
- Effectively it allows training several neural networks without any significant computational overhead.



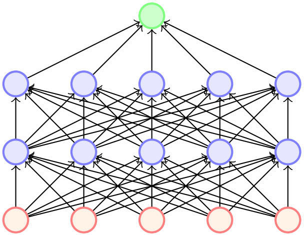
- Dropout is a technique which addresses both these issues.
- Effectively it allows training several neural networks without any significant computational overhead.
- Also gives an efficient approximate way of combining exponentially many different neural networks.

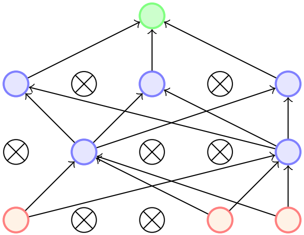


- Dropout refers to dropping out units

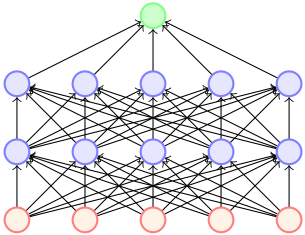


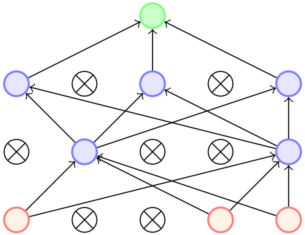
- Dropout refers to dropping out units
- Temporarily remove a node and all its incoming/outgoing connections resulting in a thinned network

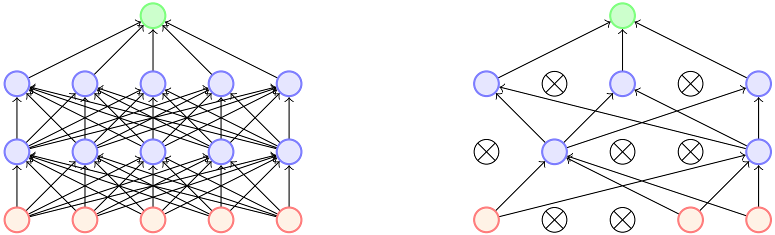




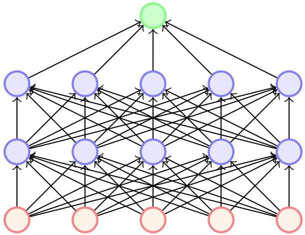
- Dropout refers to dropping out units
- Temporarily remove a node and all its incoming/outgoing connections resulting in a thinned network
- Each node is retained with a fixed probability (typically $p = 0.5$) for hidden nodes and $p = 0.8$ for visible nodes

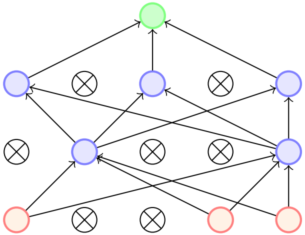




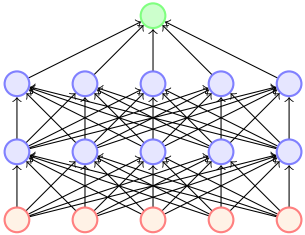


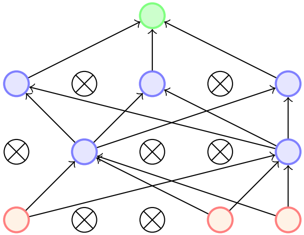
- A neural network with n nodes can be seen as a collection of 2^n possible thinned networks



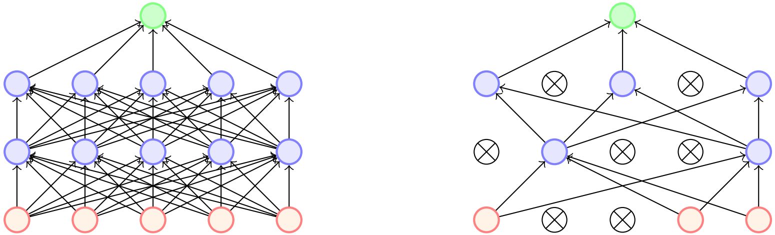


- A neural network with n nodes can be seen as a collection of 2^n possible thinned networks
- The weights in these networks are shared

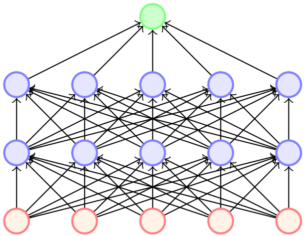


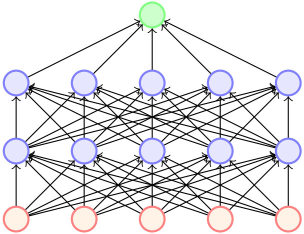


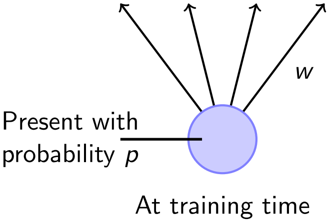
- A neural network with n nodes can be seen as a collection of 2^n possible thinned networks
- The weights in these networks are shared
- For each training instance, a different thinned network is sampled and trained



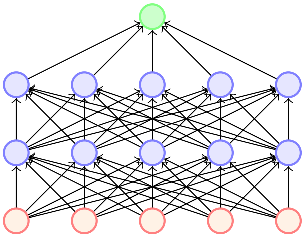
- A neural network with n nodes can be seen as a collection of 2^n possible thinned networks
- The weights in these networks are shared
- For each training instance, a different thinned network is sampled and trained
- Each thinned network gets trained rarely (or even never) but the parameter sharing ensures that no model has untrained or poorly trained parameters

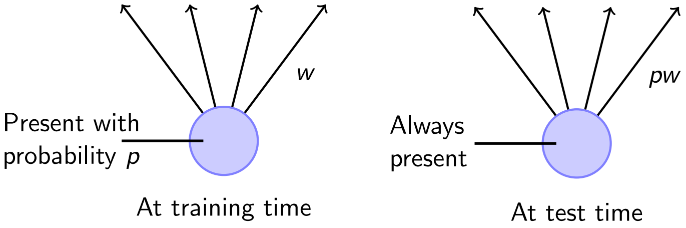




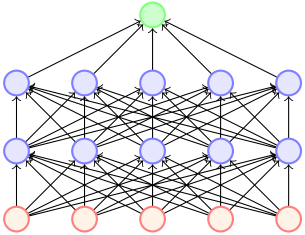


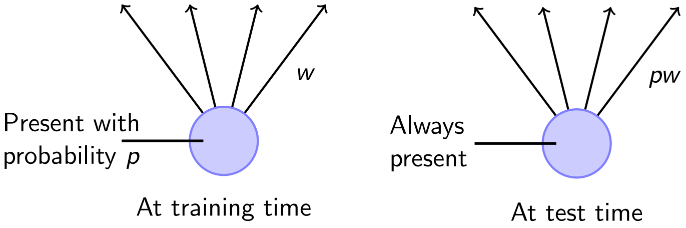
- What happens at test time?



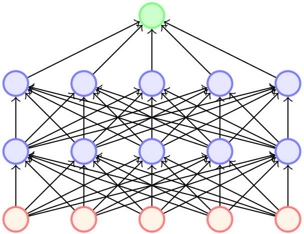


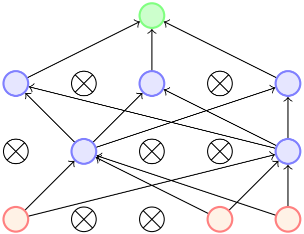
- What happens at test time?
- Impossible to aggregate the outputs of 2^n thinned networks.

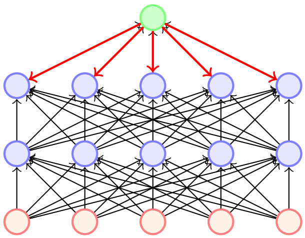
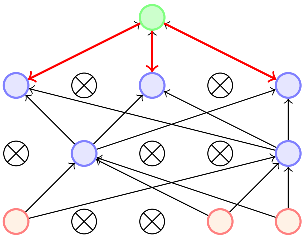




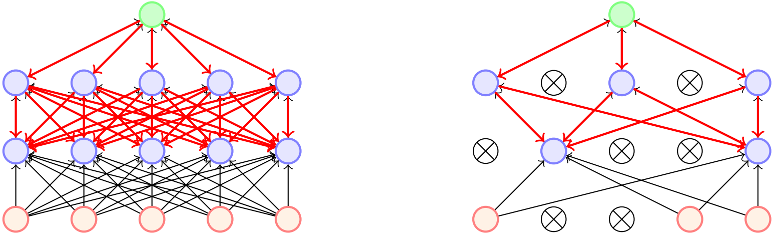
- What happens at test time?
- Impossible to aggregate the outputs of 2^n thinned networks.
- Instead we use the full Neural Network and scale the output of each node by the fraction of times it was on during training.



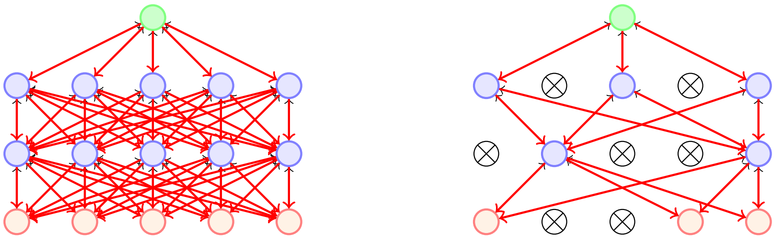


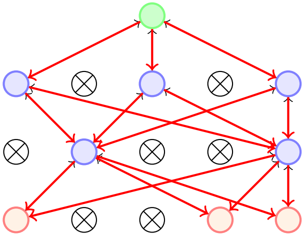
- How do you do backpropagation in such a noisy network which changes for each training instance (or batch)



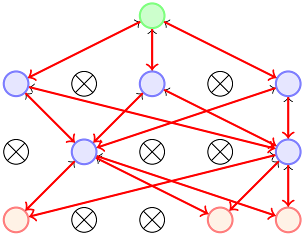
- How do you do backpropagation in such a noisy network which changes for each training instance (or batch)
- Simple: we only backpropagate over the paths which are active and only update those weights which are active in the current thinned network



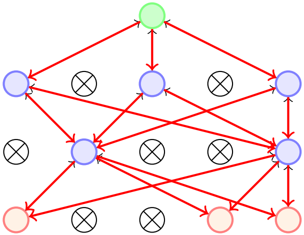
- How do you do backpropagation in such a noisy network which changes for each training instance (or batch)
- Simple: we only backpropagate over the paths which are active and only update those weights which are active in the current thinned network



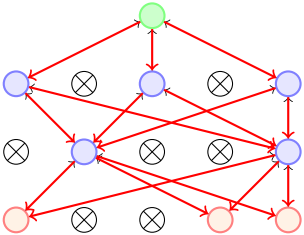
- Dropout essentially applies a masking noise to the hidden units



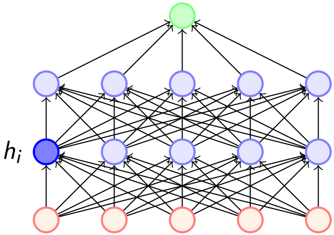
- Dropout essentially applies a masking noise to the hidden units
- Prevents hidden units from co-adapting

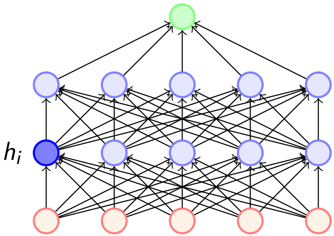


- Dropout essentially applies a masking noise to the hidden units
- Prevents hidden units from co-adapting
- Essentially a hidden unit cannot rely too much on other units as they may get dropped out any time

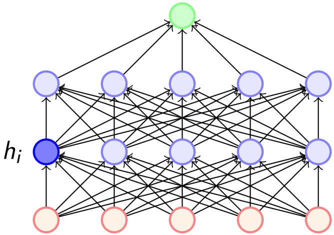


- Dropout essentially applies a masking noise to the hidden units
- Prevents hidden units from co-adapting
- Essentially a hidden unit cannot rely too much on other units as they may get dropped out any time
- Each hidden unit has to learn to be more robust to these random dropouts

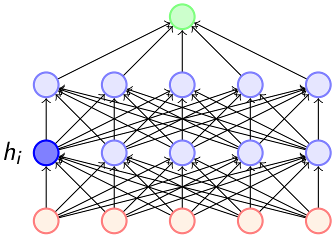

h_i



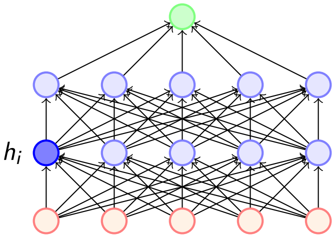
- Here is an example of how dropout helps in ensuring redundancy and robustness



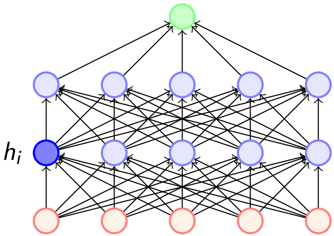
- Here is an example of how dropout helps in ensuring redundancy and robustness
- Suppose h_i learns to detect a face by firing on detecting a nose



- Here is an example of how dropout helps in ensuring redundancy and robustness
- Suppose h_i learns to detect a face by firing on detecting a nose
- Dropping h_i then corresponds to erasing the information that a noise exists



- Here is an example of how dropout helps in ensuring redundancy and robustness
- Suppose h_i learns to detect a face by firing on detecting a nose
- Dropping h_i then corresponds to erasing the information that a noise exists
- The model should then learn another h_i which redundantly encodes the presence of a nose



- Here is an example of how dropout helps in ensuring redundancy and robustness
- Suppose h_i learns to detect a face by firing on detecting a nose
- Dropping h_i then corresponds to erasing the information that a noise exists
- The model should then learn another h_i which redundantly encodes the presence of a nose
- Or the model should learn to detect the face using other features

Recap

- $L2$ regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout