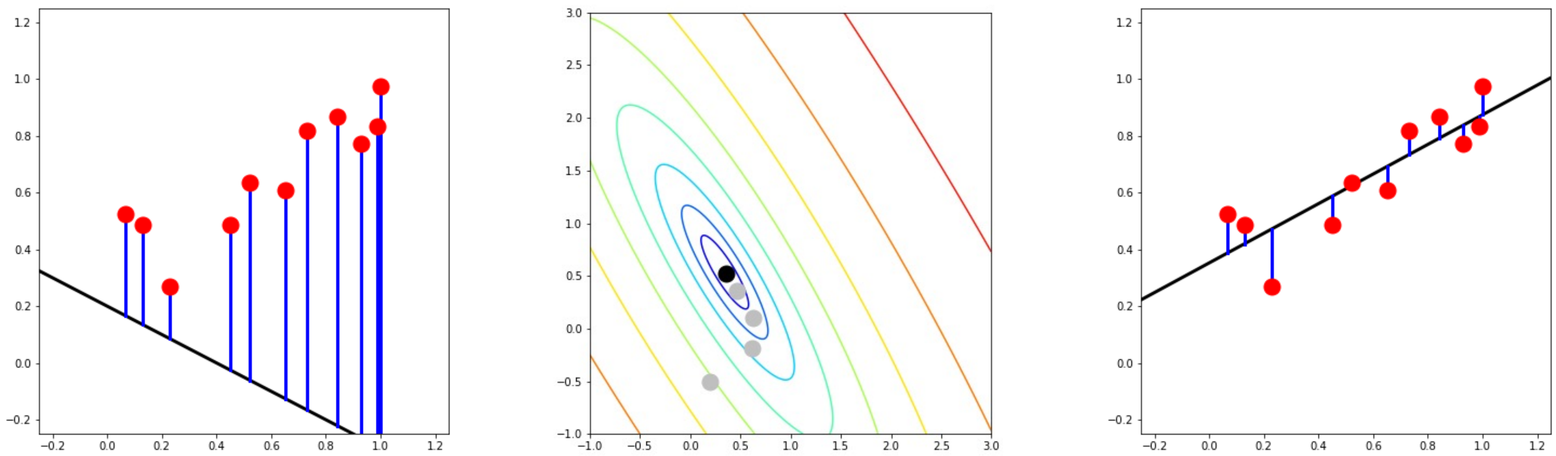
CS273A:
Linear Regression



Prof. Alexander Ihler

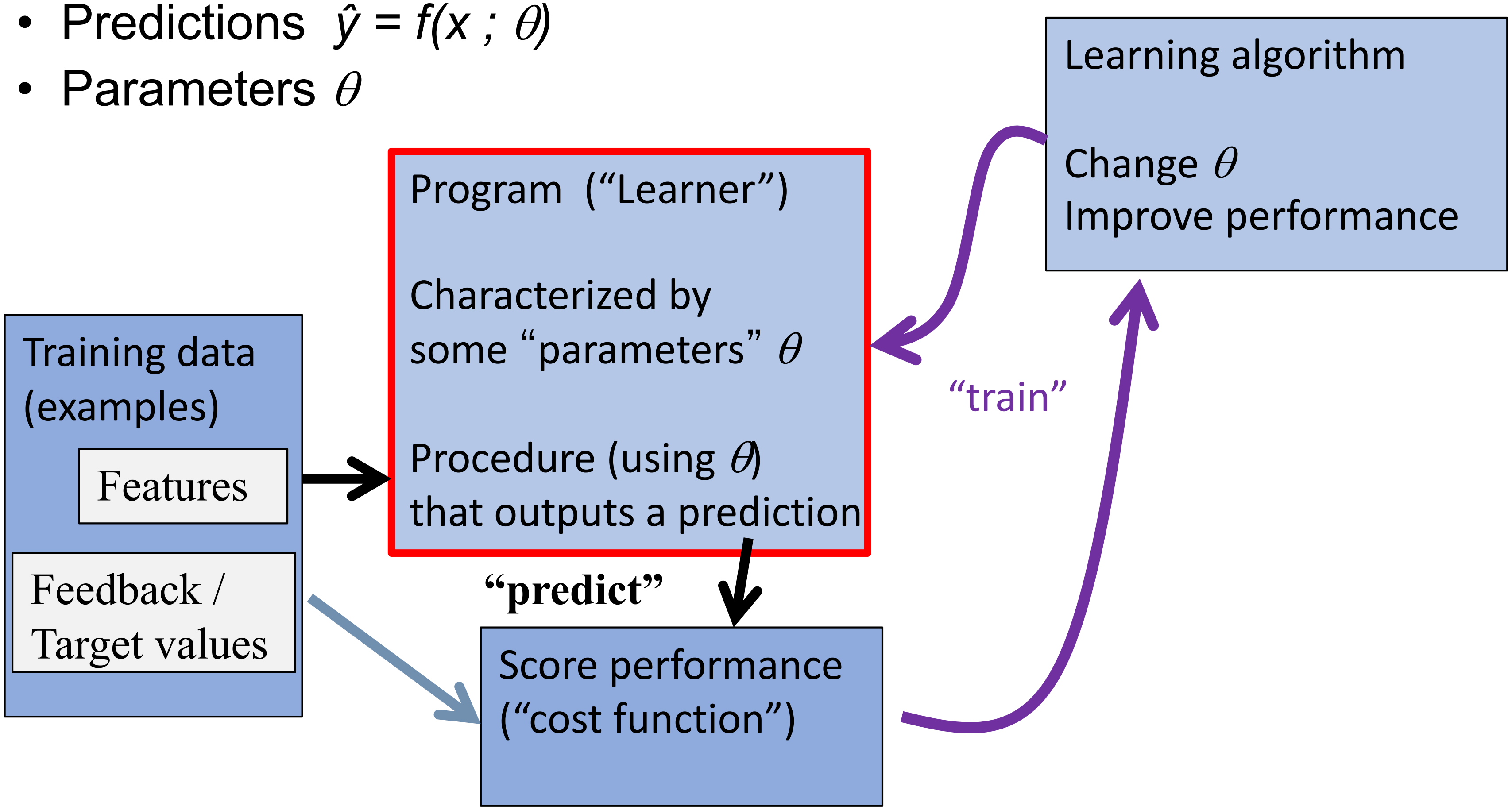
Fall 2024

Linear Regression

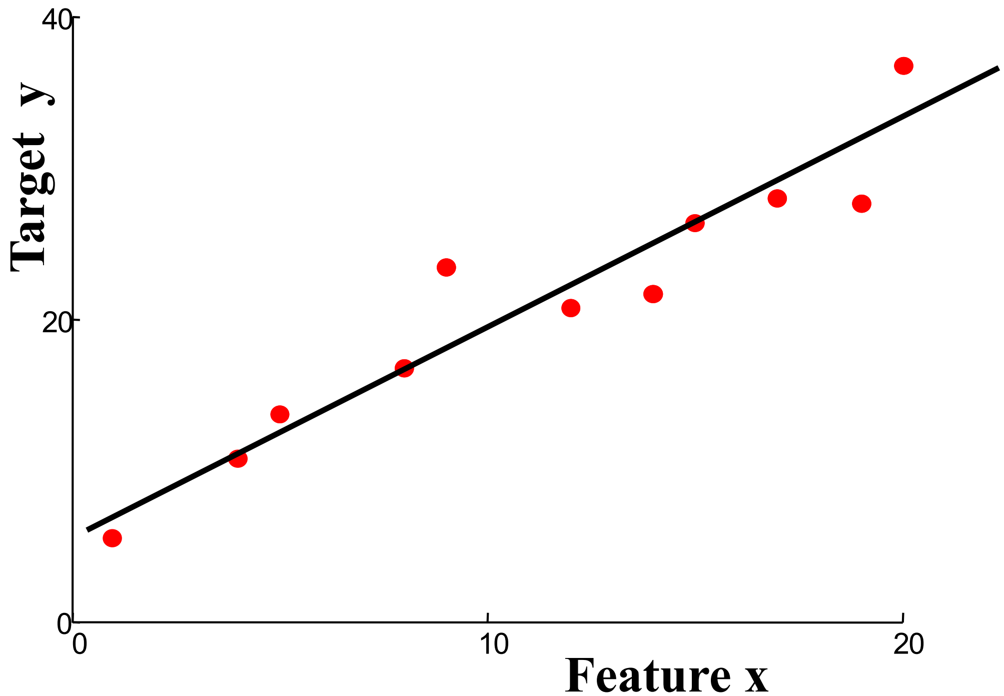
- **Linear Regression via Least Squares**
- Gradient Descent Algorithms
- Direct Minimization of Squared Error
- Regression with Non-linear Features
- Bias, Variance, & Validation
- Regularized Linear Regression

Supervised learning

- Notation
  - Features x
  - Targets y
  - Predictions $\hat{y} = f(x\,;\,\theta)$
  - Parameters θ



Linear regression



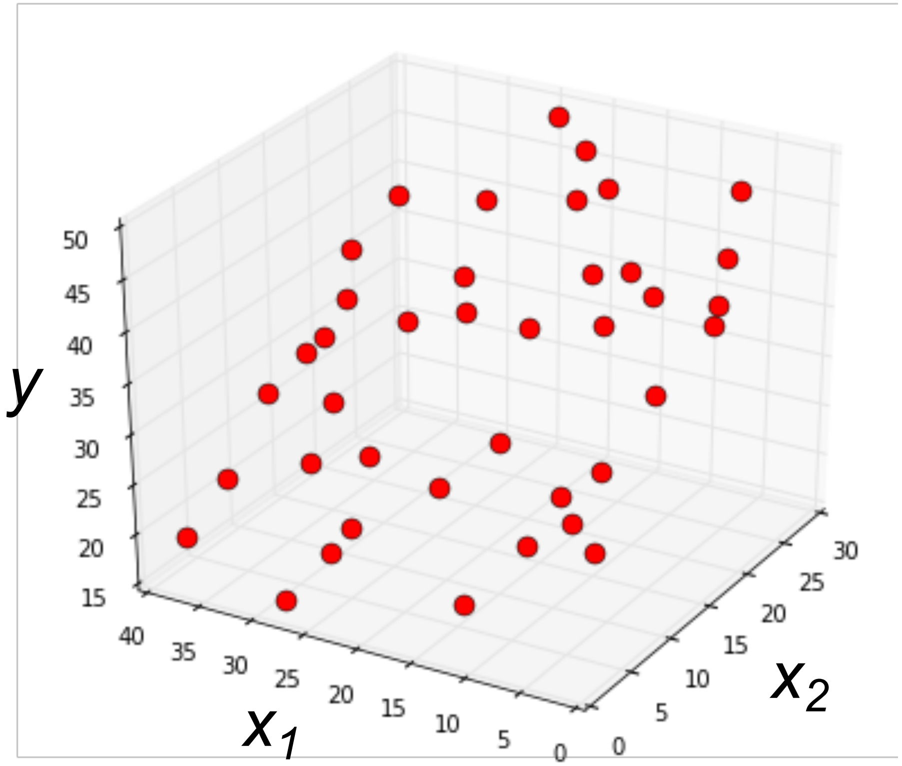
"Predictor":
Evaluate line:

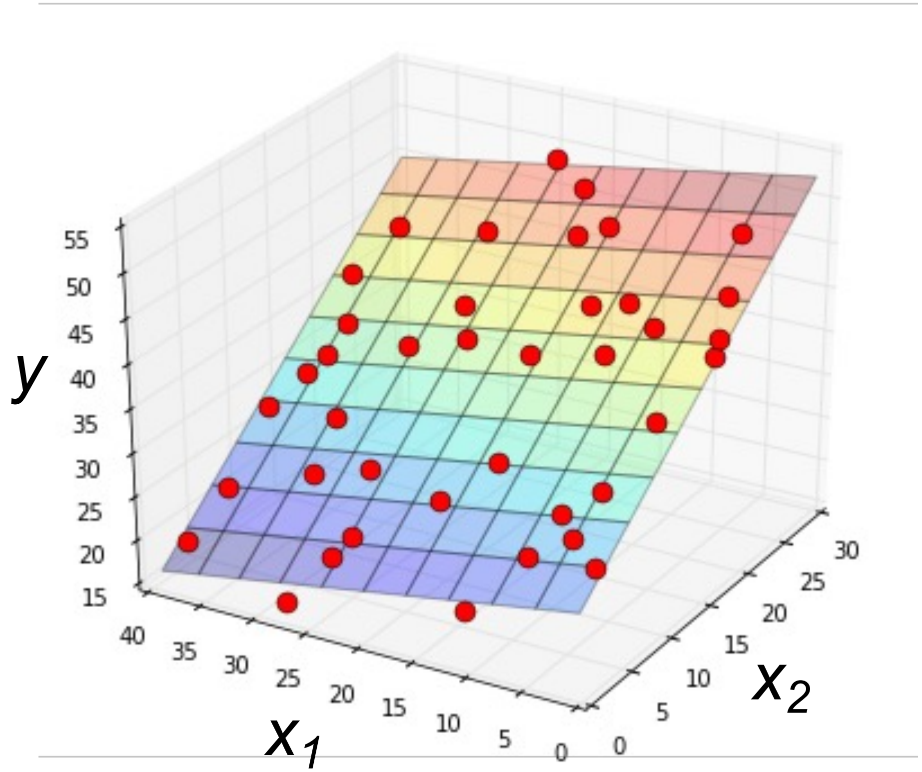
$$r = \theta_0 + \theta_1 x_1$$

return r

- Define form of function f(x) explicitly
- Find a good f(x) within that family

More dimensions?





"Predictor":
Evaluate linear response:

$$r = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

return r

Notation

$$\hat{y}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots$$

Define "feature" $x_0 = 1$ (constant)

Then

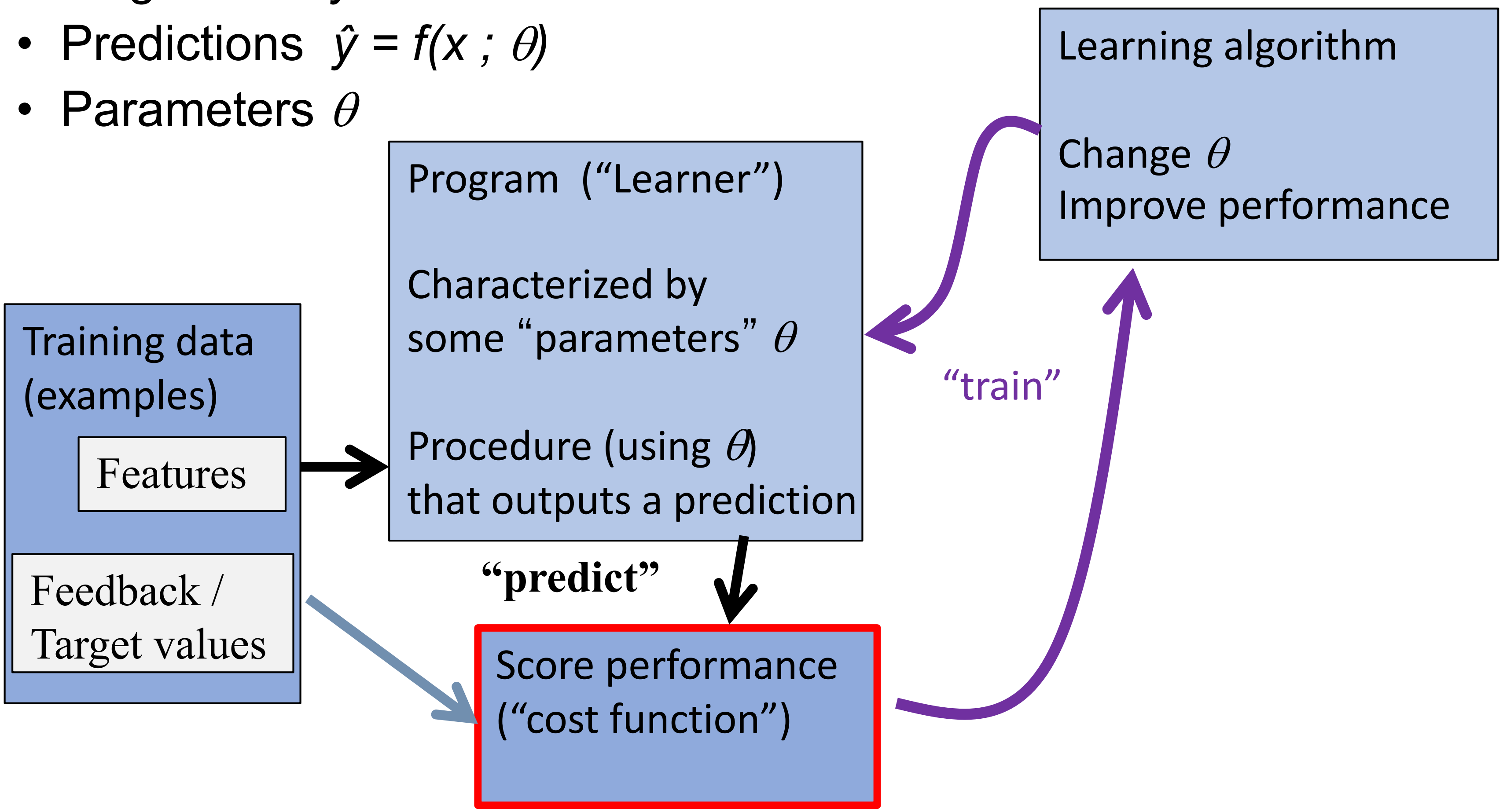
$$\hat{y}(x) = \theta \, x^T$$

$$\underline{\theta} = [\theta_0, \dots, \theta_n]$$

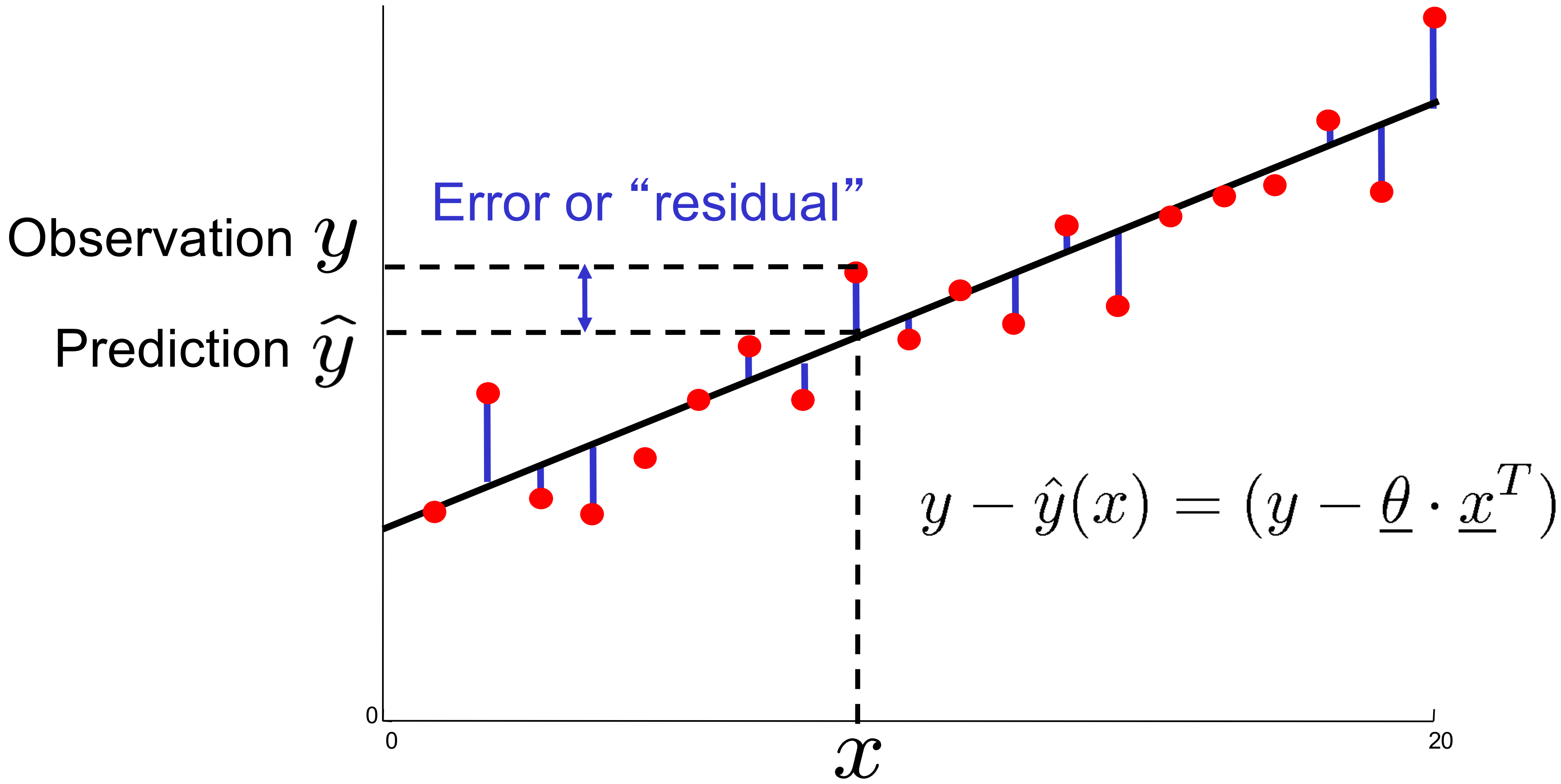
$$\underline{x} = [1, x_1, \dots, x_n]$$

Supervised learning

- Notation
  - Features x
  - Targets y
  - Predictions $\hat{y} = f(x\,;\,\theta)$
  - Parameters θ



Measuring error



Mean squared error

- How can we quantify the error?

$$\text{MSE},\ J(\underline{\theta}) = \frac{1}{m}\sum_j (y^{(j)} - \hat{y}(x^{(j)}))^2$$

$$= \frac{1}{m}\sum_j (y^{(j)} - \underline{\theta} \cdot \underline{x}^{(j)T})^2$$

- Could choose something else, of course…
  - Computationally convenient (more later)
  - Measures the variance of the residuals
  - Corresponds to likelihood under Gaussian model of "noise"

$$\mathcal{N}(y\ ;\ \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

MSE cost function

$$\text{MSE},\ J(\underline{\theta}) = \frac{1}{m}\sum_j (y^{(j)} - \hat{y}(x^{(j)}))^2$$

$$= \frac{1}{m}\sum_j (y^{(j)} - \underline{\theta}\cdot \underline{x}^{(j)T})^2$$

- Rewrite using matrix form

$$\underline{\theta} = [\theta_0, \ldots, \theta_n]$$

$$\underline{y} = \left[y^{(1)} \ldots, y^{(m)}\right]^T$$

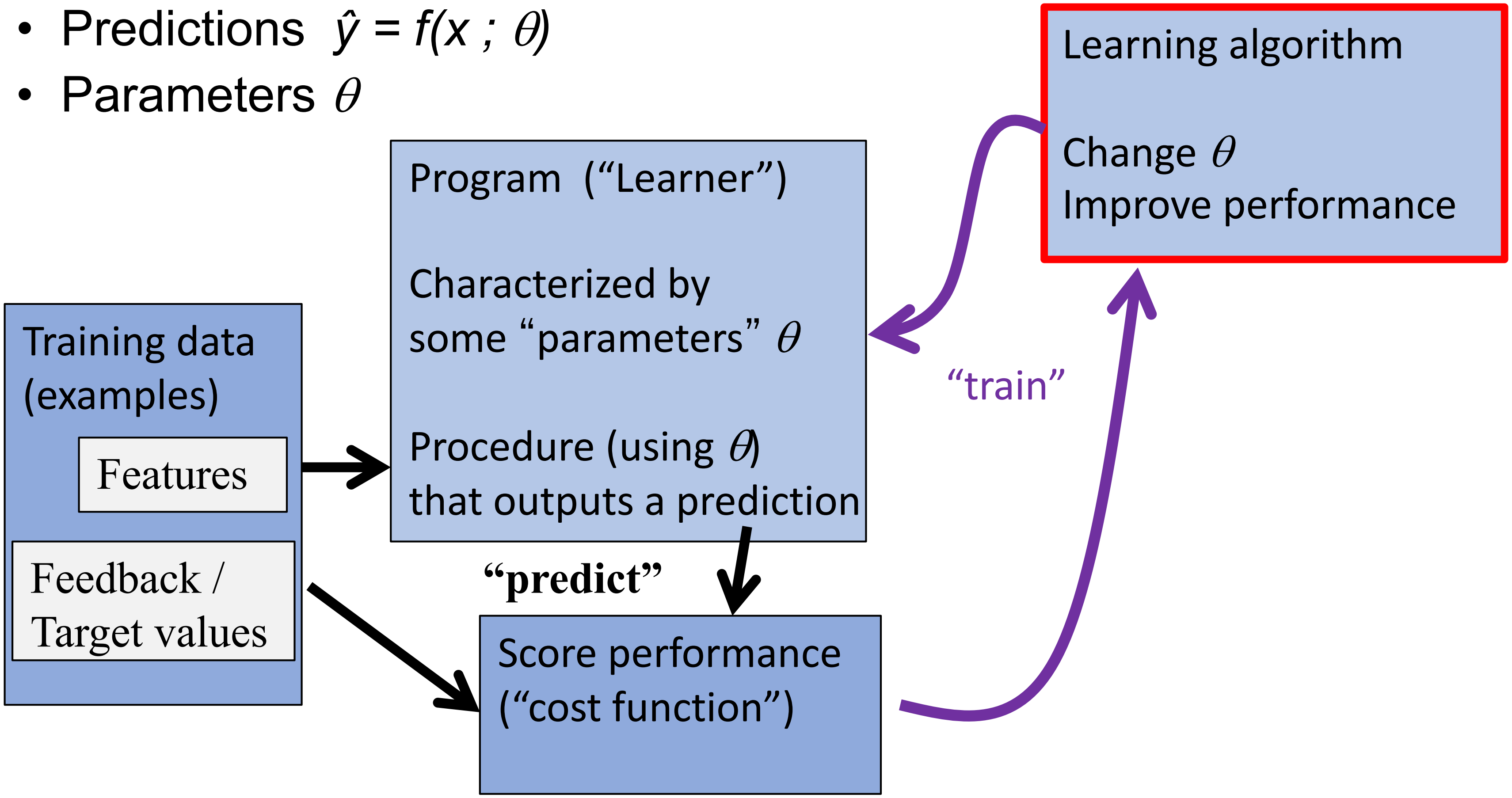
$$\underline{X} = \begin{bmatrix} x_0^{(1)} & \ldots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_0^{(m)} & \ldots & x_n^{(m)} \end{bmatrix}$$

$$J(\underline{\theta}) = \frac{1}{m}(\underline{y}^T - \underline{\theta}\,\underline{X}^T)\cdot(\underline{y}^T - \underline{\theta}\,\underline{X}^T)^T$$

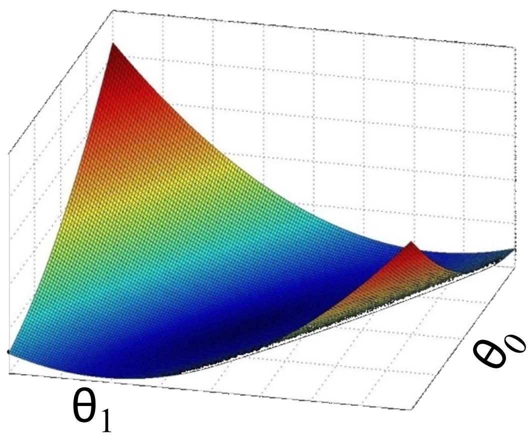
```
# Python / NumPy:
e = Y – X.dot( theta.T );
J = e.T.dot( e ) / m  # = np.mean( e ** 2 )
```

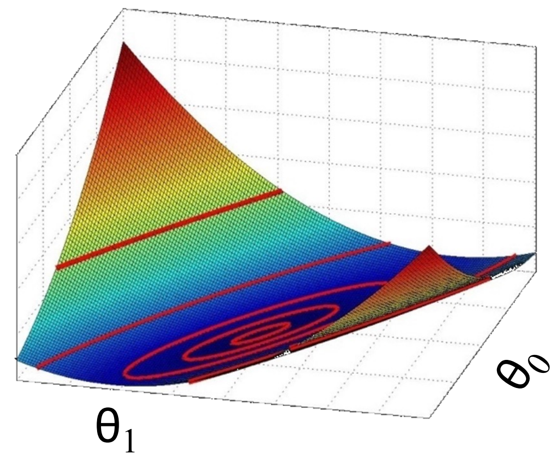
Supervised learning

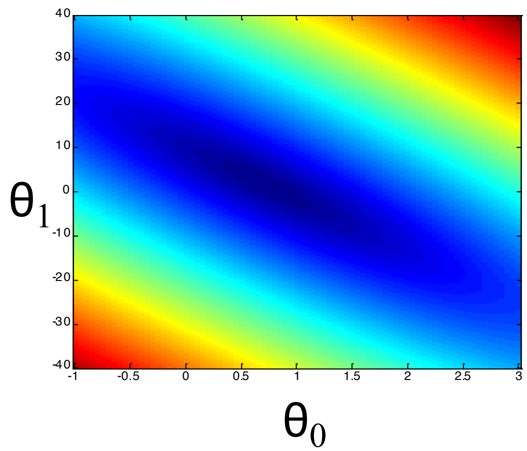
- Notation
  - Features x
  - Targets y
  - Predictions $\hat{y} = f(x\ ;\ \theta)$
  - Parameters θ

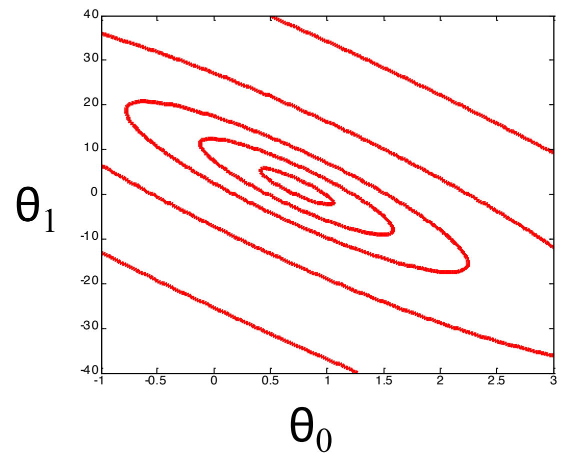


Visualizing the cost function



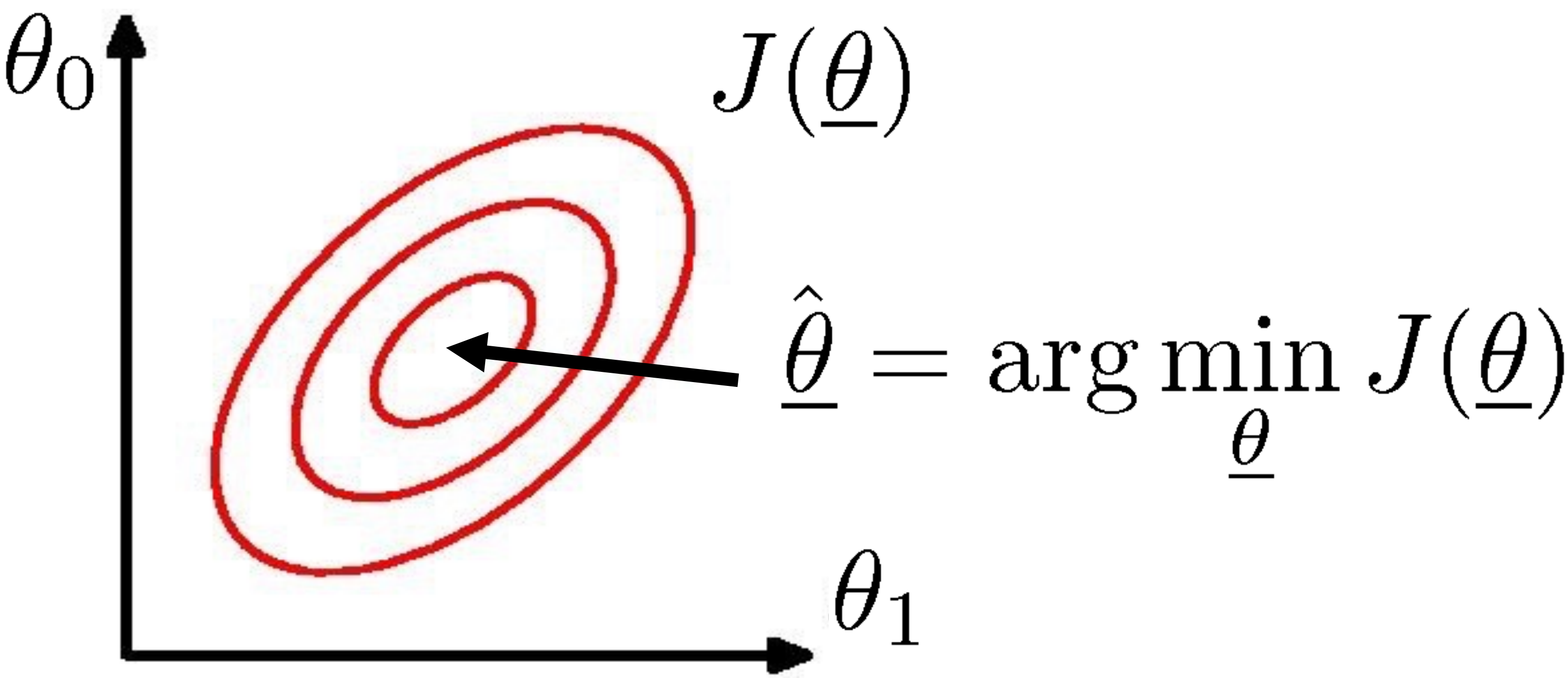






Finding good parameters

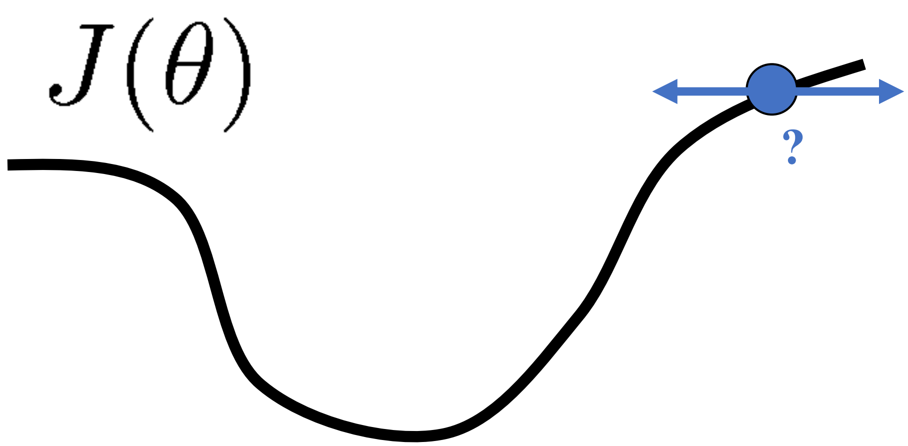
- Want to find parameters which minimize our error…
- Think of a cost "surface": error residual for that θ …



Linear Regression

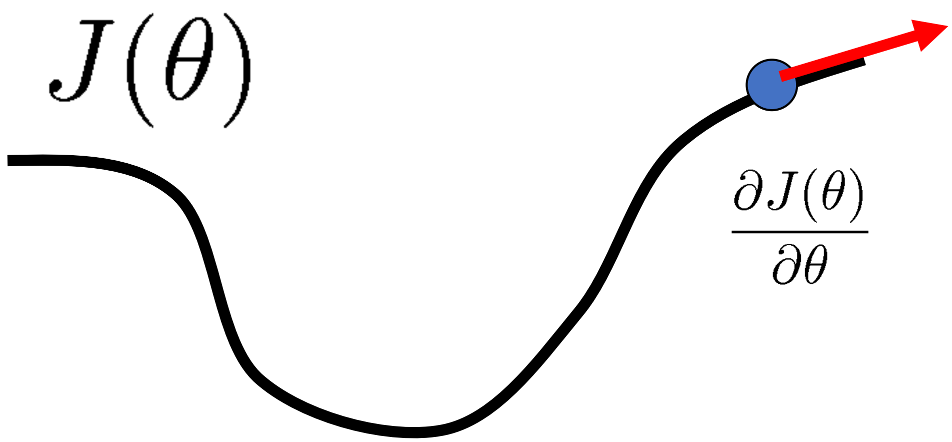
- Linear Regression via Least Squares
- Gradient Descent Algorithms
- Direct Minimization of Squared Error
- Regression with Non-linear Features
- Bias, Variance, & Validation
- Regularized Linear Regression

Gradient descent



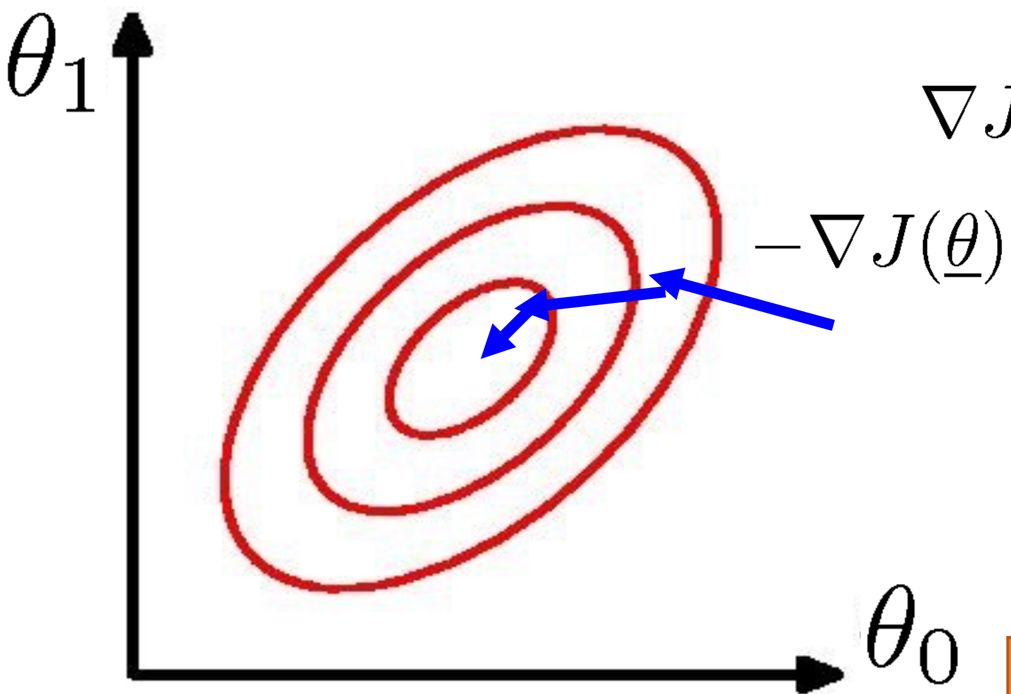
- How to change θ to improve J(θ)?
- Choose a direction in which J(θ) is decreasing

Gradient descent



- How to change θ to improve J(θ)?
- Choose a direction in which J(θ) is decreasing
- Derivative $\frac{\partial J(\theta)}{\partial \theta}$
- Positive => increasing
- Negative => decreasing

Gradient descent in more dimensions



- Gradient vector

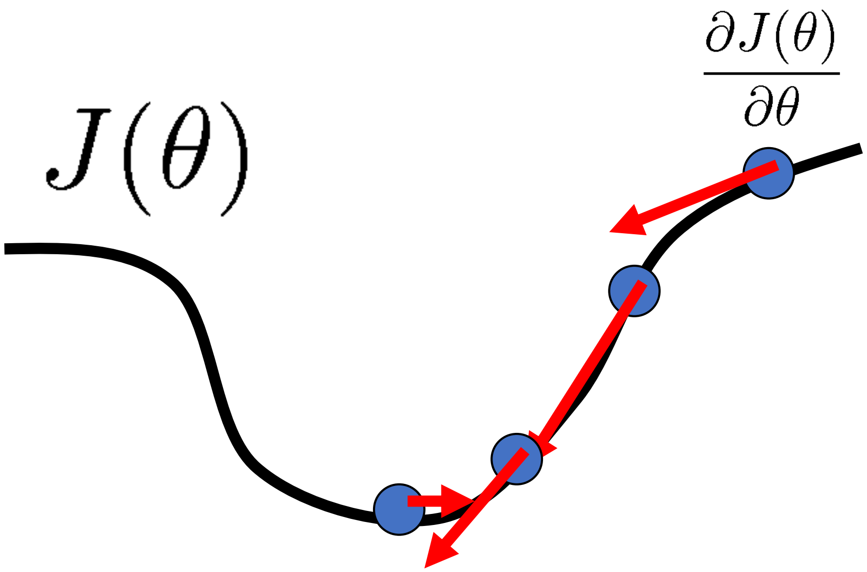
$$\nabla J(\underline{\theta}) = \left[\frac{\partial J(\underline{\theta})}{\partial \theta_0} \quad \frac{\partial J(\underline{\theta})}{\partial \theta_1} \quad \ldots \right]$$

Indicates direction of steepest ascent (negative = steepest descent)

Gradient descent

- Initialization
- Step size α
  - Can change over iterations
- Gradient direction
- Stopping condition

```
Initialize θ
Do{
 θ ← θ - α∇θJ(θ)
} while (α‖∇θJ‖ > ε)
```



Gradient for the MSE

- MSE $\quad J(\underline{\theta}) = \frac{1}{m} \sum_j (y^{(j)} - \underline{\theta} \cdot \underline{x}^{(j)T})^2$

- ∇ J = ?

$$J(\underline{\theta}) = \frac{1}{m} \sum_j (\overbrace{y^{(j)} - \theta_0 \underline{x}_0^{(j)} - \theta_1 \underline{x}_1^{(j)} - \ldots}^{e_j(\theta)})^2$$

$$\frac{\partial J}{\partial \theta_0} = \frac{\partial}{\partial \theta_0} \frac{1}{m} \sum_j (\; e_j(\theta) \;)^2$$

$$= \frac{1}{m} \sum_j \quad \frac{\partial}{\partial \theta_0} (\; e_j(\theta) \;)^2$$

$$= \frac{1}{m} \sum_j \; 2 e_j(\theta) \; \frac{\partial}{\partial \theta_0} e_j(\theta)$$

$$\frac{\partial}{\partial \theta_0} e_j(\theta) = \cancelto{0}{\frac{\partial}{\partial \theta_0} y^{(j)}} - \frac{\partial}{\partial \theta_0} \theta_0 x_0^{(j)} - \cancelto{0}{\frac{\partial}{\partial \theta_0} \theta_1 x_1^{(j)}} - \ldots$$

$$= -x_0^{(j)}$$

Gradient for the MSE

- MSE $$J(\underline{\theta}) = \frac{1}{m} \sum_j (y^{(j)} - \underline{\theta} \cdot \underline{x}^{(j)^T})^2$$

- ∇ J = ?
$$J(\underline{\theta}) = \frac{1}{m} \sum_j (\underbrace{y^{(j)} - \theta_0 \underline{x}_0^{(j)} - \theta_1 \underline{x}_1^{(j)} - \ldots}_{e_j(\theta)})^2$$

$$\nabla J(\underline{\theta}) = \left[\quad \frac{\partial J}{\partial \theta_0} \qquad\qquad \frac{\partial J}{\partial \theta_1} \qquad \ldots \right]$$

$$= \left[\underbrace{\frac{2}{m} \sum_j -e_j(\theta) x_0^{(j)}} \quad \underbrace{\frac{2}{m} \sum_j -e_j(\theta) x_1^{(j)}} \qquad \ldots \right]$$

Gradient descent

- Initialization
- Step size α
  - Can change over iterations
- Gradient direction
- Stopping condition

```
Initialize θ
Do{
 θ ← θ - α∇θJ(θ)
} while (α‖∇θJ‖ > ε)
```

$$J(\underline{\theta}) = \frac{1}{m} \sum_j (y^{(j)} - \underline{\theta} \cdot \underline{x}^{(j)^T})^2$$

$$\nabla J(\underline{\theta}) = -\frac{2}{m} \sum_j (y^{(j)} \underbrace{- \underline{\theta} \cdot \underline{x}^{(j)^T}}_{\text{Error magnitude \& direction for datum j}}) \cdot [\underbrace{x_0^{(j)} x_1^{(j)}}_{\text{Sensitivity to each parameter}} \ldots]$$

Derivative of MSE

- Rewrite using matrix form

$$\nabla J(\underline{\theta}) = -\frac{2}{m} \sum_j \underbrace{(y^{(j)} - \underline{\theta} \cdot \underline{x}^{(j)T})}_{\text{Error magnitude \& direction for datum j}} \cdot [\underbrace{x_0^{(j)} x_1^{(j)}}_{\text{Sensitivity to each } \theta_i} \ldots]$$

$$\underline{\theta} = [\theta_0, \ldots, \theta_n]$$

$$\underline{y} = \left[y^{(1)} \ldots, y^{(m)}\right]^T$$

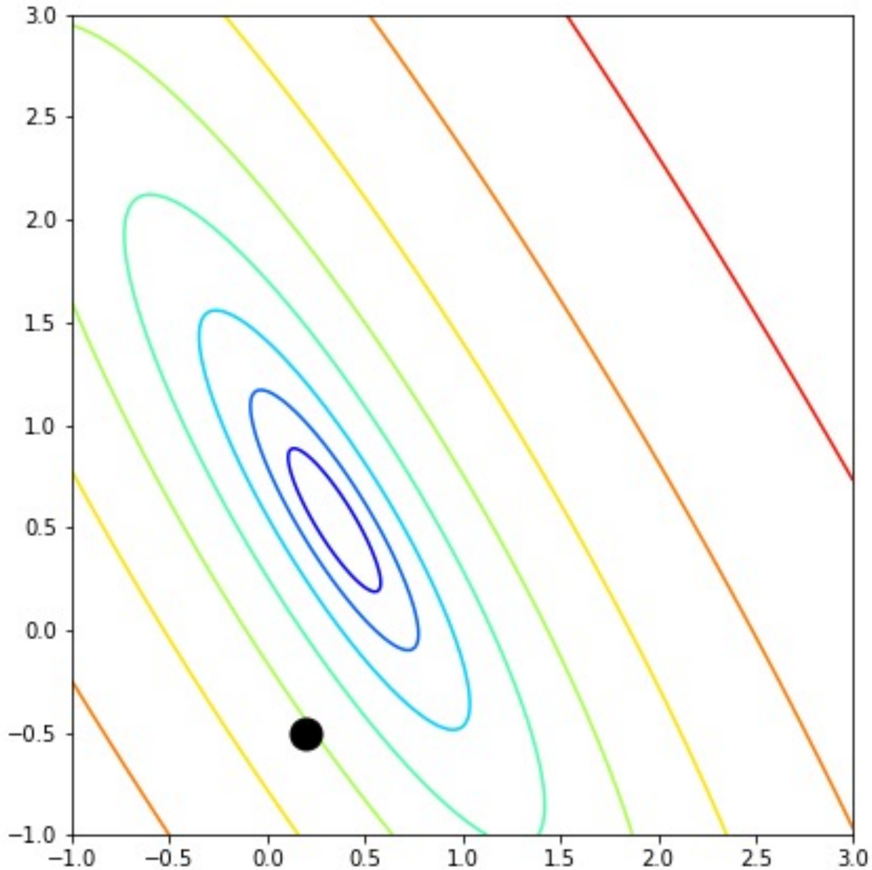
$$\nabla J(\underline{\theta}) = -\frac{2}{m} (\underline{y}^T - \underline{\theta X}^T) \cdot \underline{X}$$

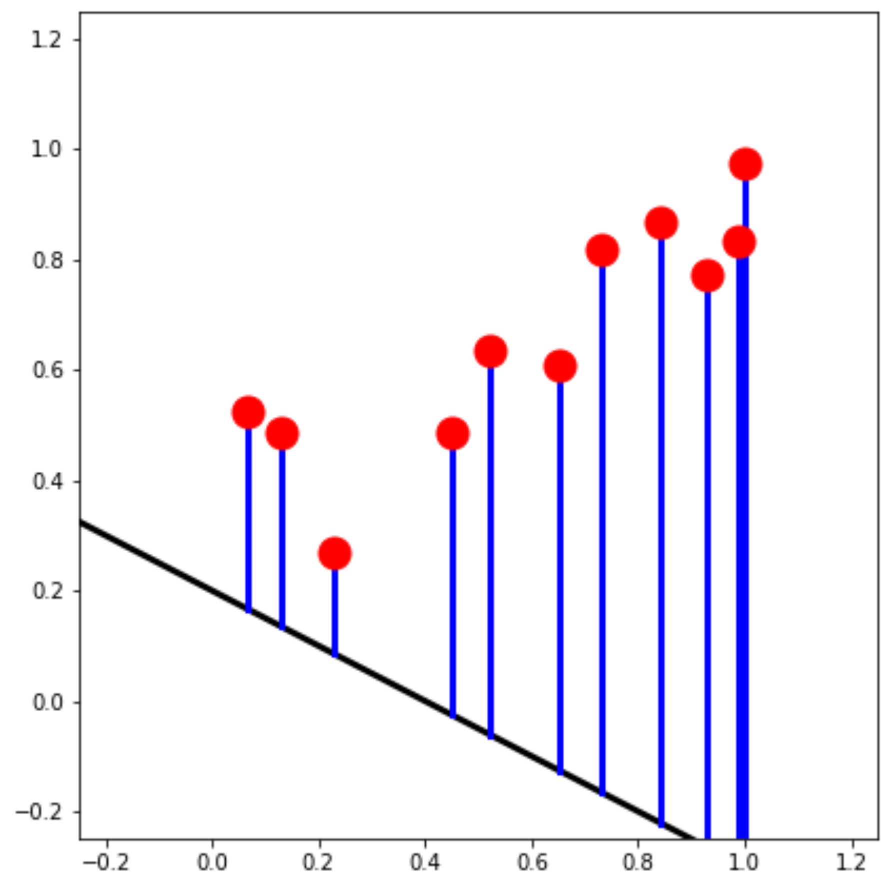
$$\underline{X} = \begin{bmatrix} x_0^{(1)} & \ldots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_0^{(m)} & \ldots & x_n^{(m)} \end{bmatrix}$$

```
e = Y – X.dot( theta.T ) # error residual
DJ = - e.dot(X) * 2.0/m  # compute the gradient
theta -= alpha * DJ       # take a step
```

Gradient descent on cost function

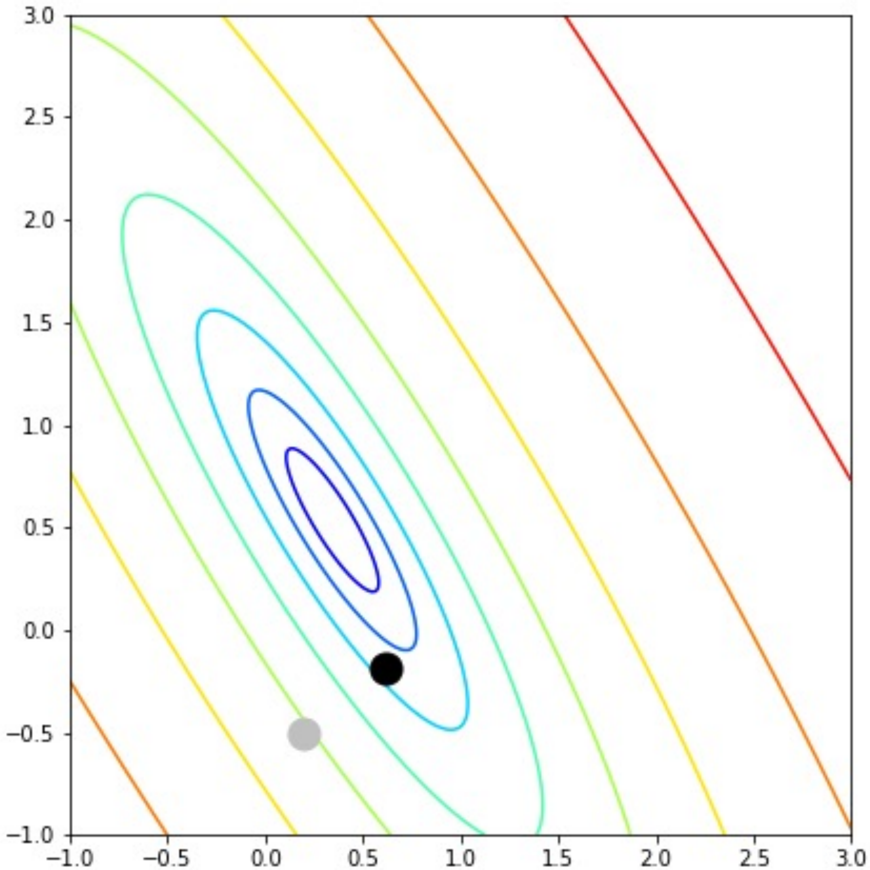
Initialization

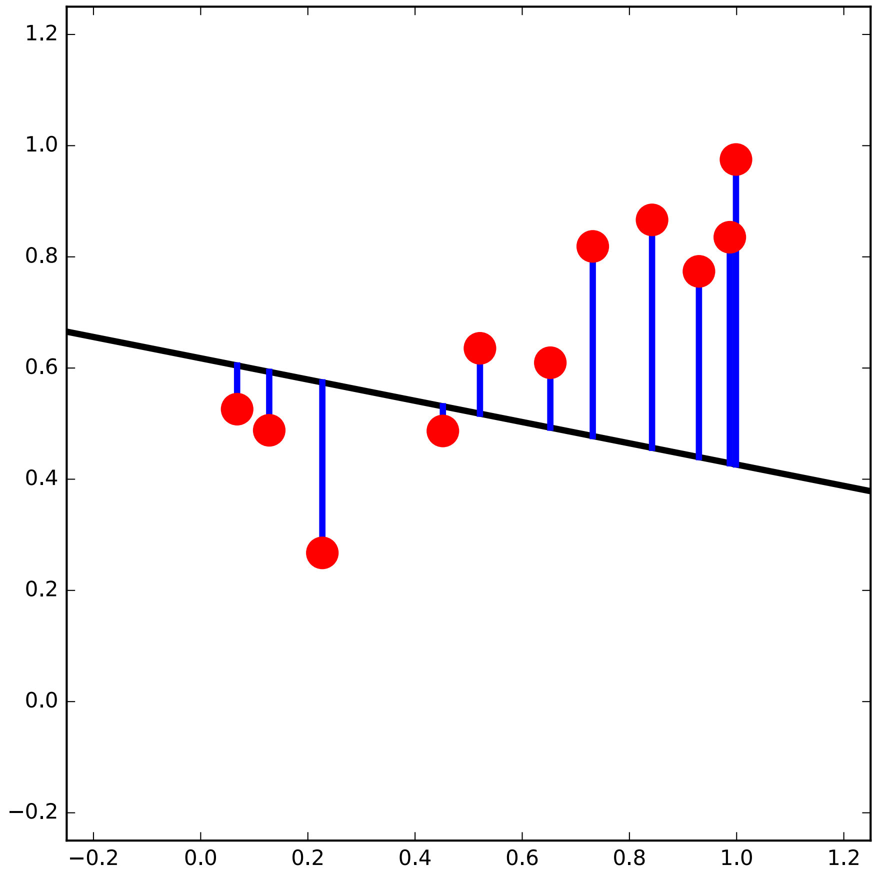




Gradient descent on cost function

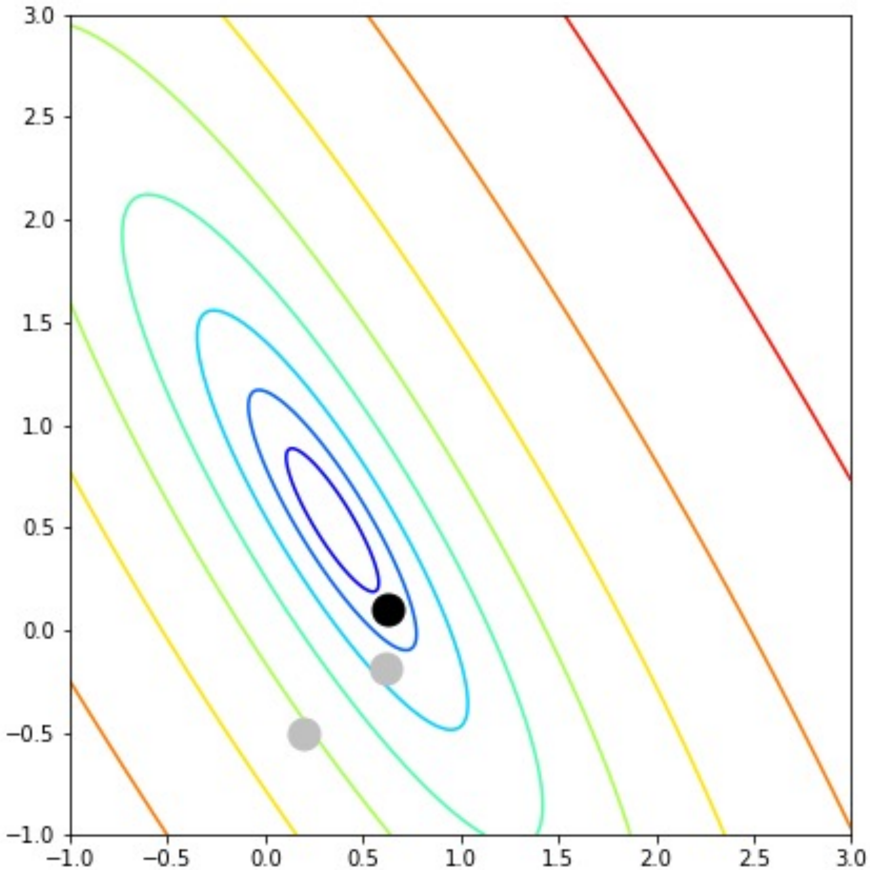
Iteration 1

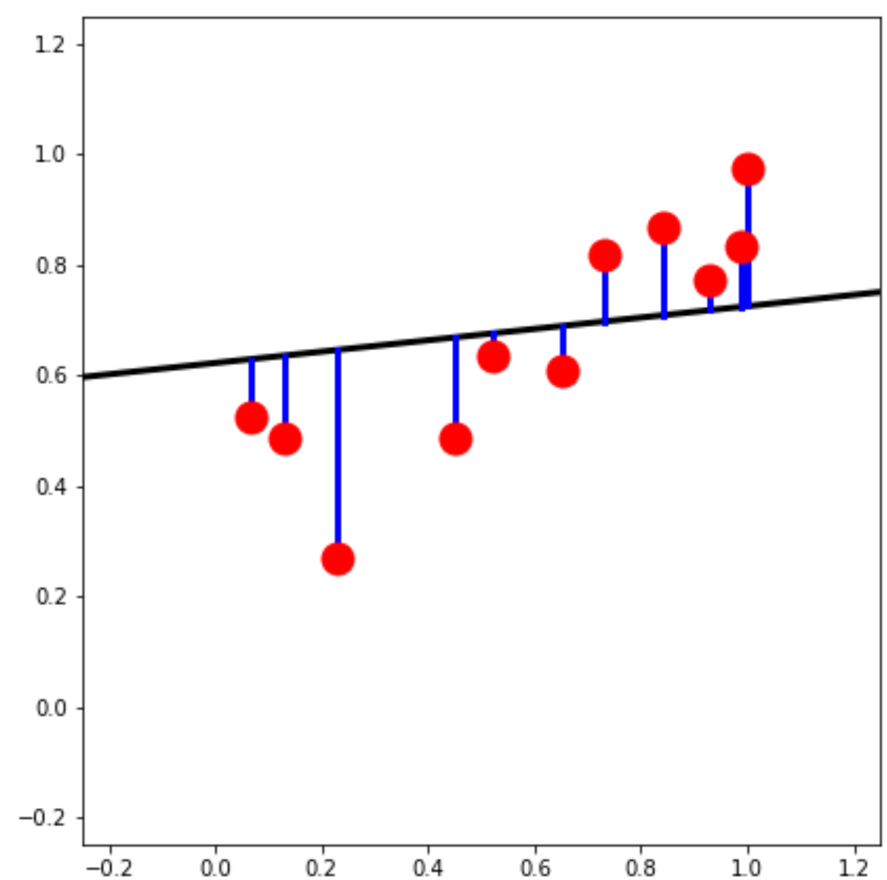




Gradient descent on cost function

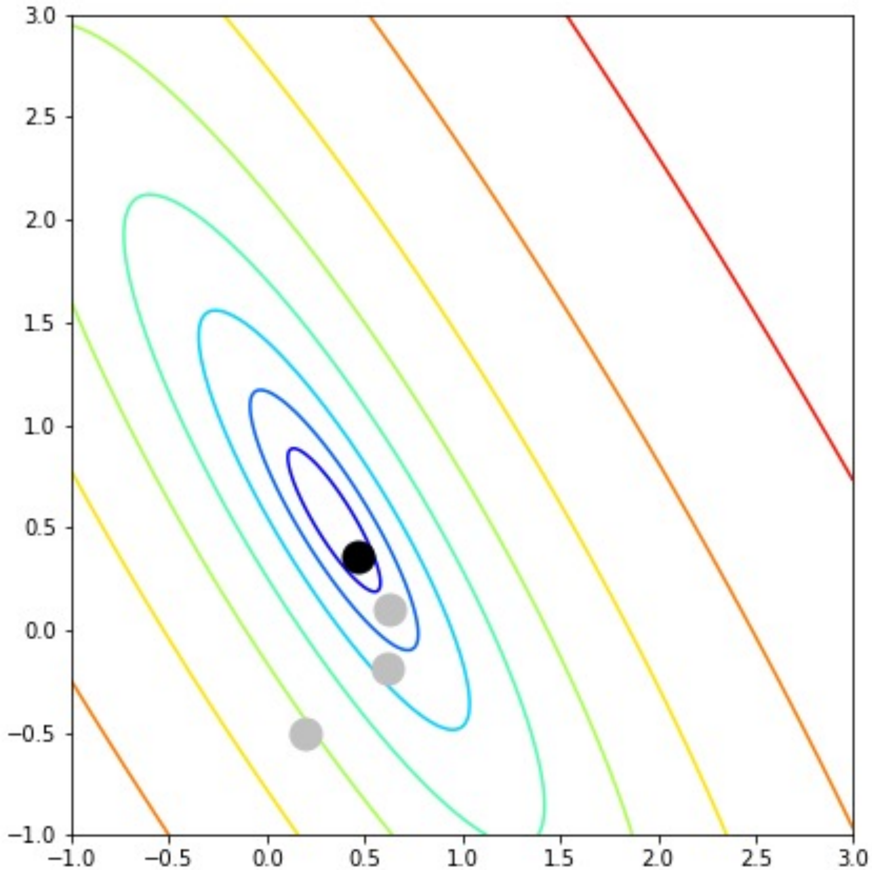
Iteration 10

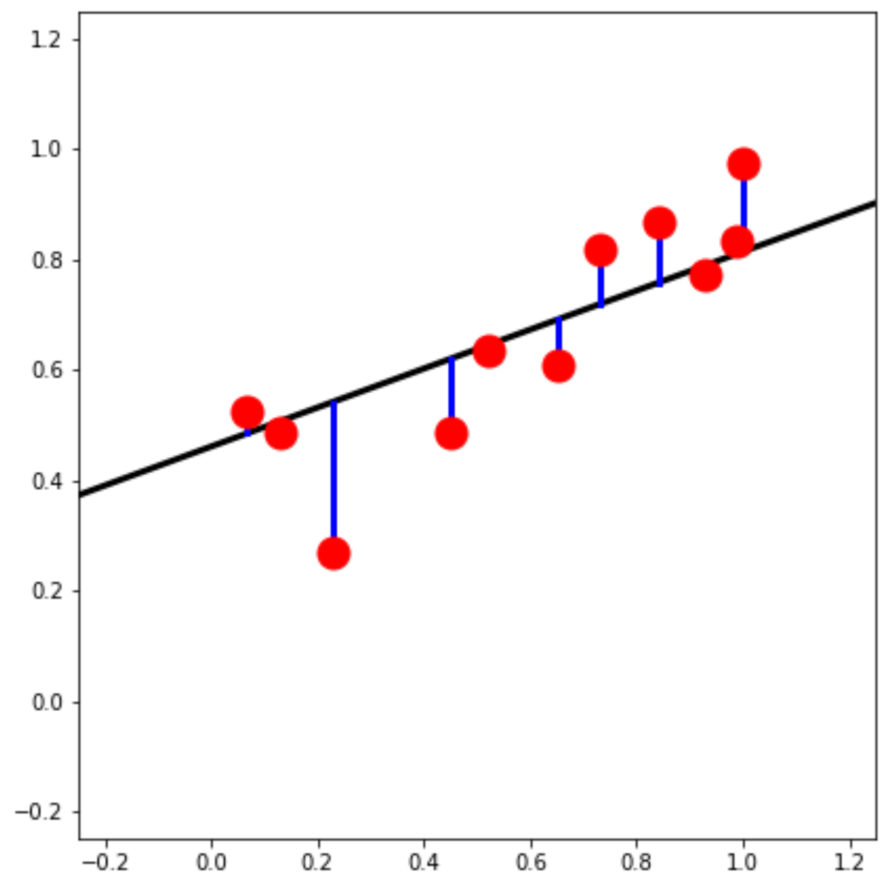




Gradient descent on cost function

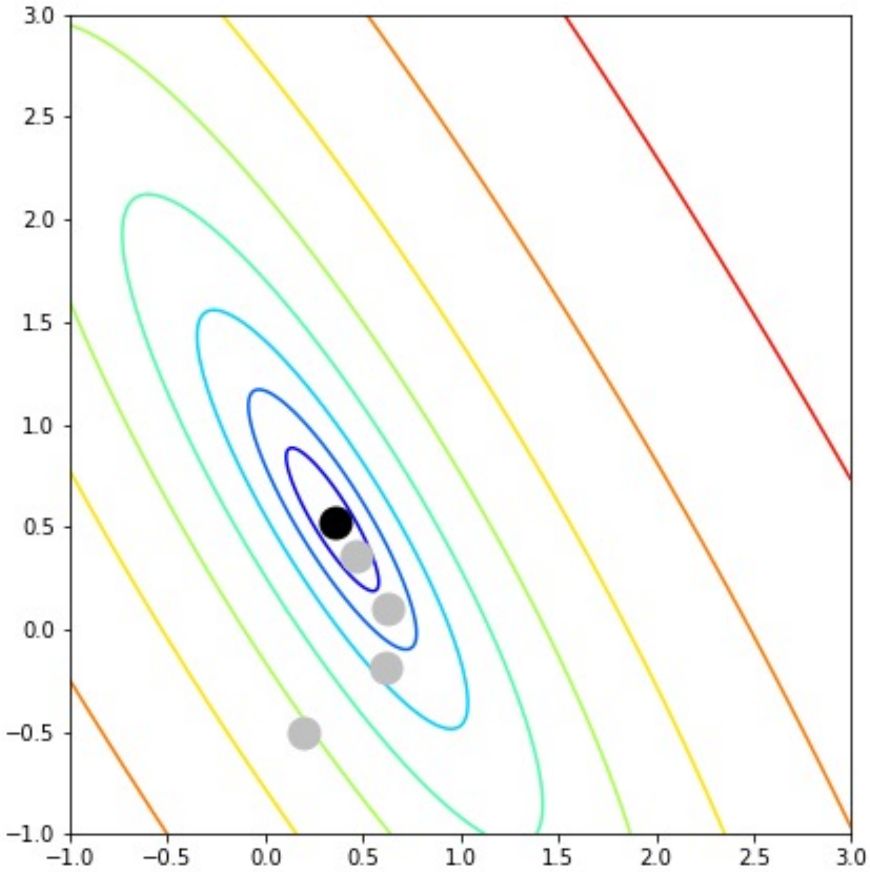
Iteration 30

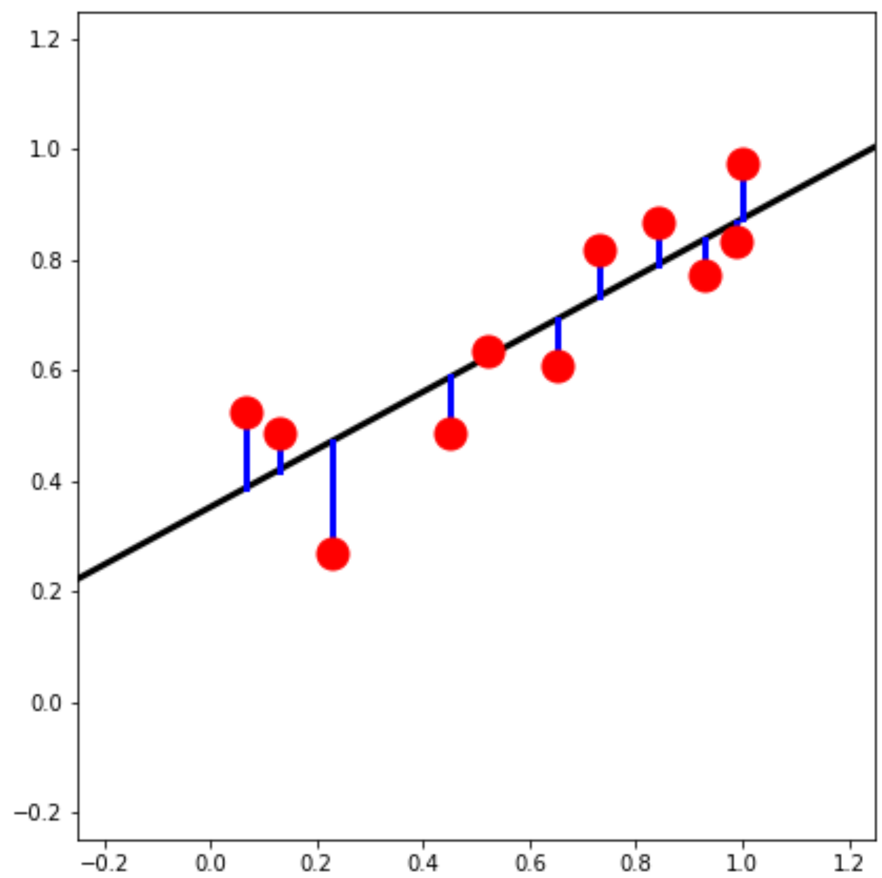




Gradient descent on cost function

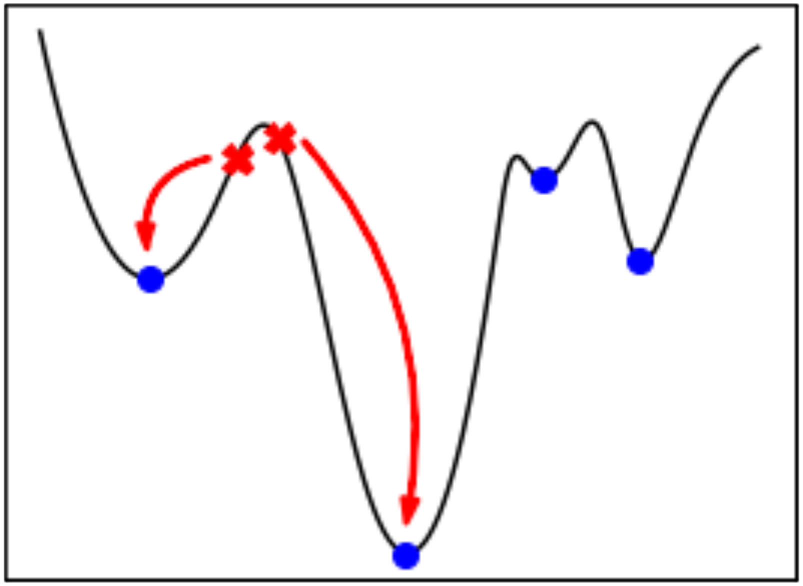
Iteration 90



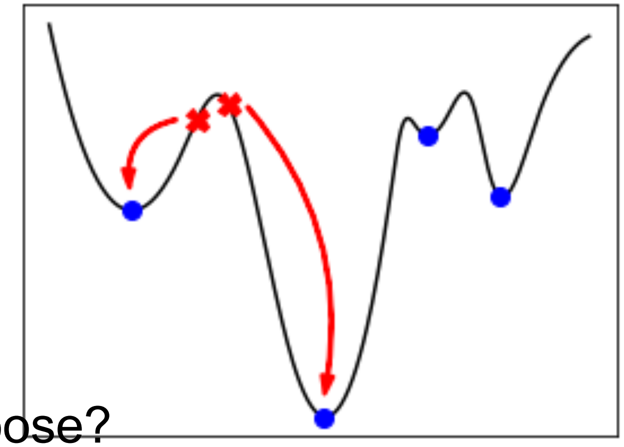


Comments on gradient descent

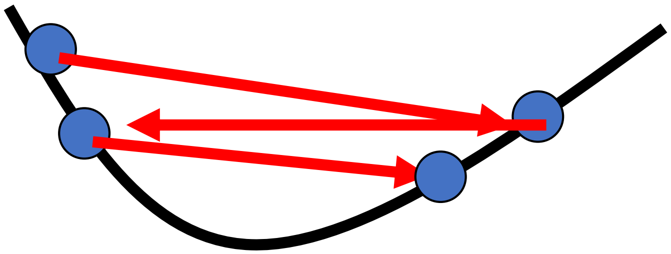
- Very general algorithm
  - We'll see it many times
- Local minima
  - Sensitive to starting point

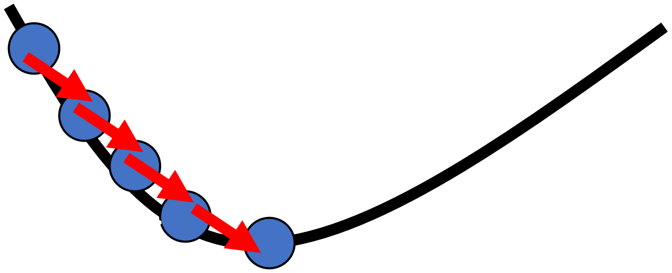


Comments on gradient descent

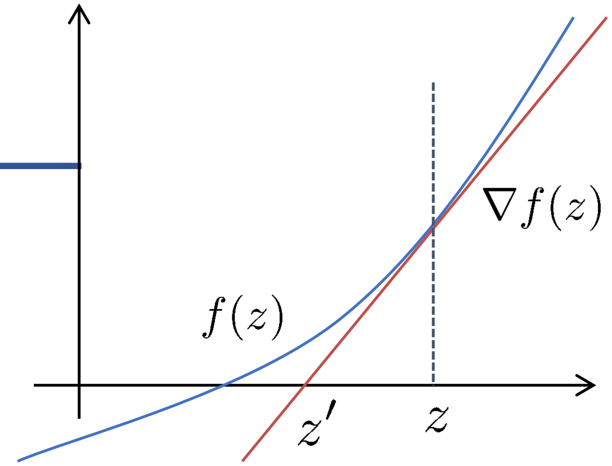


- Very general algorithm
  - We'll see it many times
- Local minima
  - Sensitive to starting point
- Step size
  - Too large? Too small? Automatic ways to choose?
  - May want step size to decrease with iteration
  - Common choices:
    - Fixed
    - Linear: C/(iteration)
    - Line search / backoff (Armijo, etc.)
    - Newton's method





Newton's method



- Want to find the roots of f(x)
    - "Root": value of x for which f(x)=0
- Initialize to *some* point x
- Compute the tangent at x & compute where it crosses x-axis

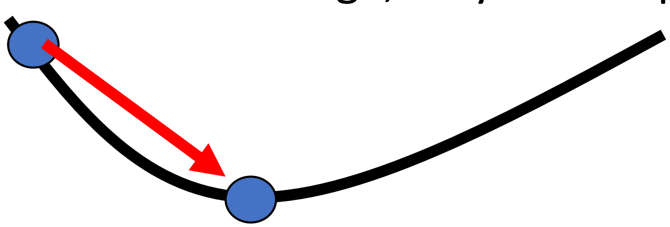
$$\nabla f(z) = \frac{0 - f(z)}{z' - z} \quad \Rightarrow \quad z' = z - \frac{f(z)}{\nabla f(z)}$$

- Optimization: find roots of $\nabla J(\theta)$

$$\nabla\nabla J(\theta) = \frac{0 - \nabla J(\theta)}{\theta' - \theta} \quad \Rightarrow \quad \theta' = \theta - \frac{\nabla J(\theta)}{\nabla\nabla J(\theta)}$$

("Step size" λ = $1/\nabla\nabla J$; inverse curvature)

- Does not always converge; sometimes unstable
- If converges, usually very fast
- Works well for smooth, non-pathological functions, locally quadratic
- For n large, may be computationally hard: $O(n^2)$ storage, $O(n^3)$ time



(Multivariate:
$\nabla J(\theta)$ = gradient vector
$\nabla^2 J(\theta)$ = matrix of 2^{nd} derivatives
$a/b = a\, b^{-1}$, matrix inverse)

Stochastic / Online gradient descent

- MSE

$$J(\theta) = \frac{1}{m} \sum_i J^{(i)}(\theta) \qquad J^{(i)}(\theta) = \left(y^{(i)} - \theta \cdot x^{(i)\,T} \right)^2$$

- Gradient

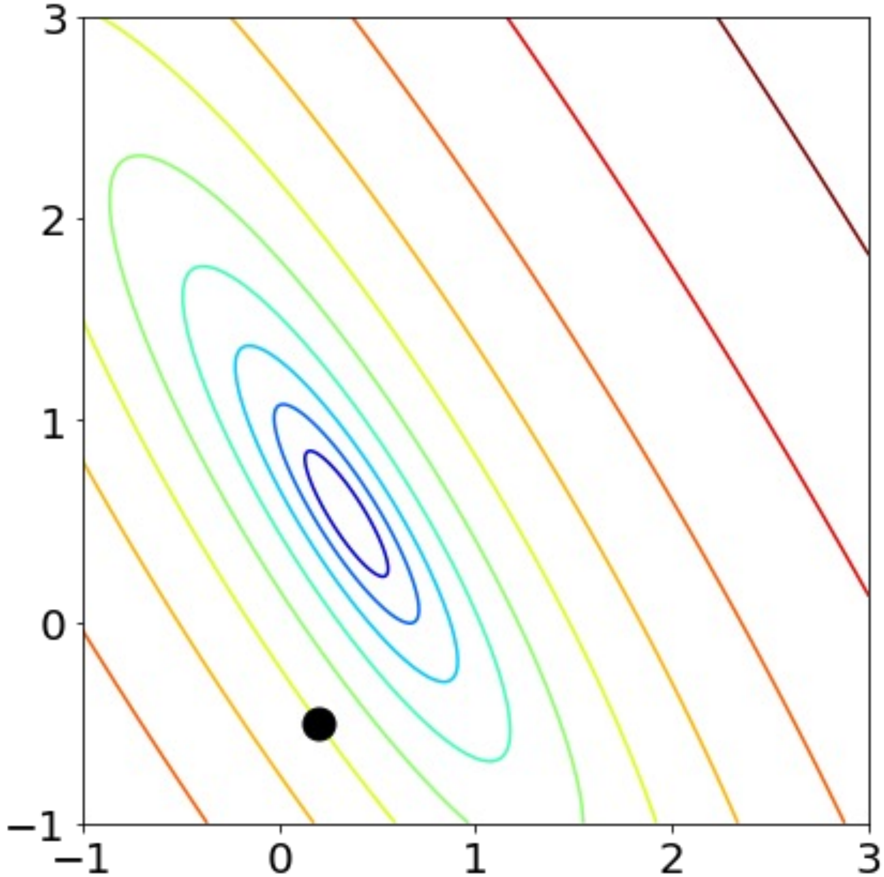
$$\nabla J(\theta) = \frac{1}{m} \sum_i \nabla J^{(i)}(\theta) \qquad \nabla J^{(i)}(\theta) = \left(y^{(i)} - \theta \cdot x^{(i)\,T} \right) [x_0^{(i)} \ x_1^{(i)} \ \ldots]$$

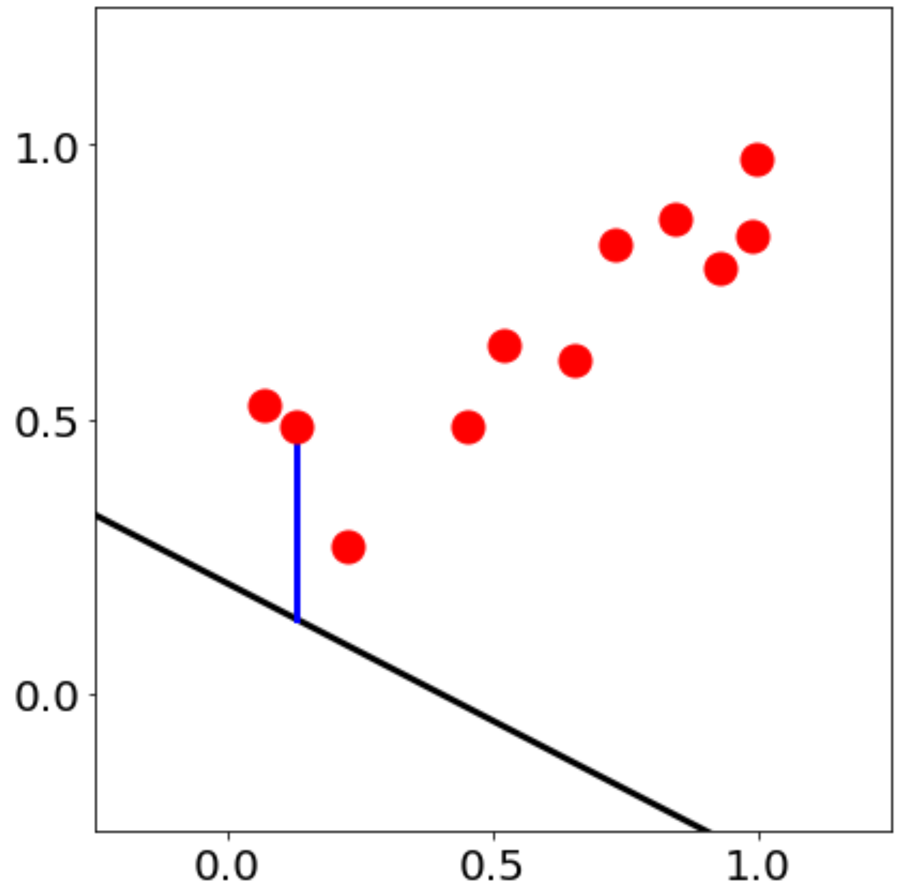
- Stochastic (or “online”) gradient descent:
  - Use updates based on individual datum j, chosen at random
  - At optima, $\mathbb{E}\big[\nabla J^{(i)}(\theta)\big] = \nabla J(\theta) = 0$
    (average over the data)

Stochastic Gradient Descent

- Initialize theta
- Select a data point & update

```
Initialize θ
Do {
 for i=1:m
   θ ← θ - α∇θJ(i)(θ)
} while (not done)
```

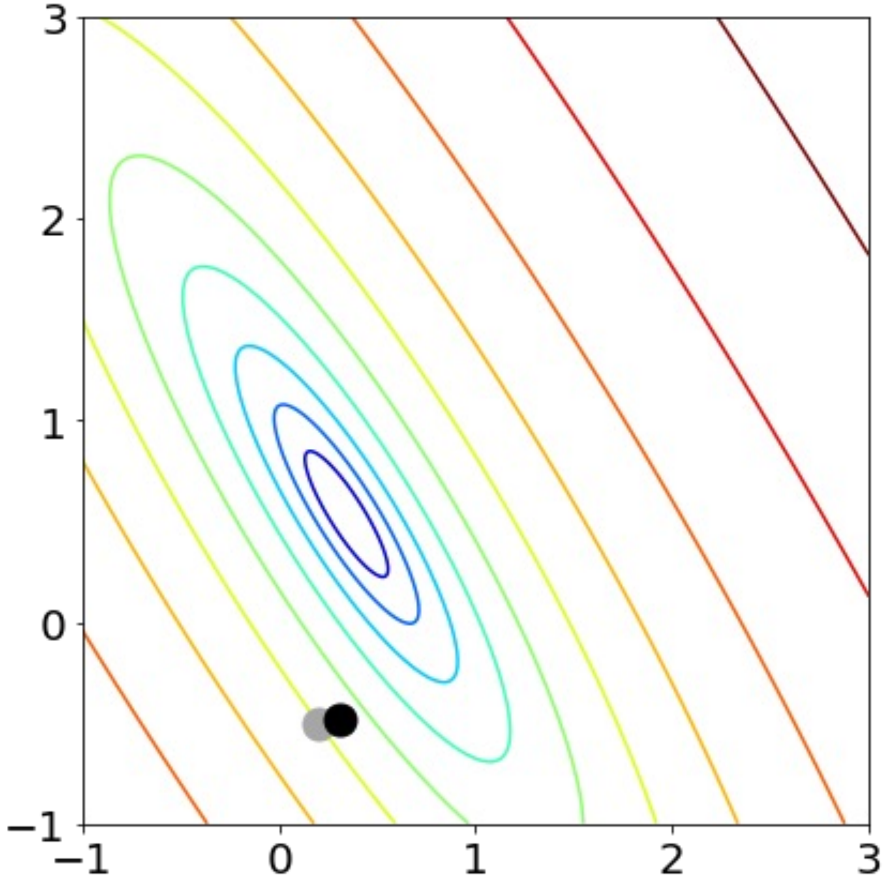


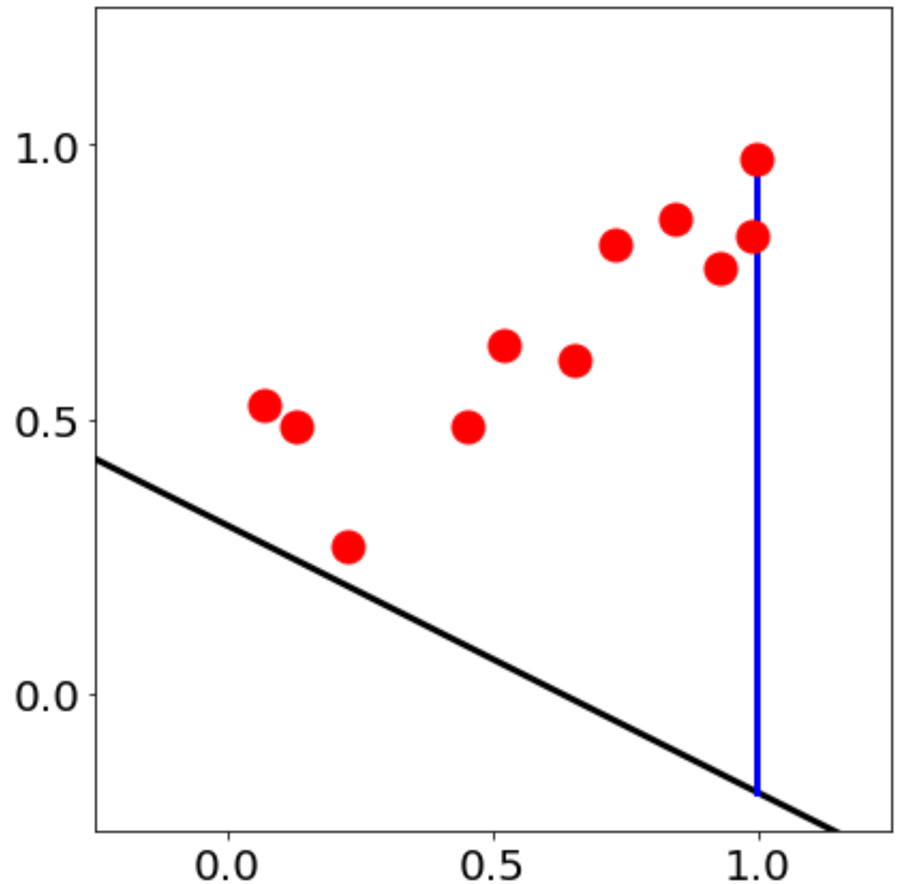


Stochastic Gradient Descent

- Initialize theta
- Select a data point & update,

```
Initialize θ
Do {
 for i=1:m
   θ ← θ - α∇θJ(i)(θ)
} while (not done)
```

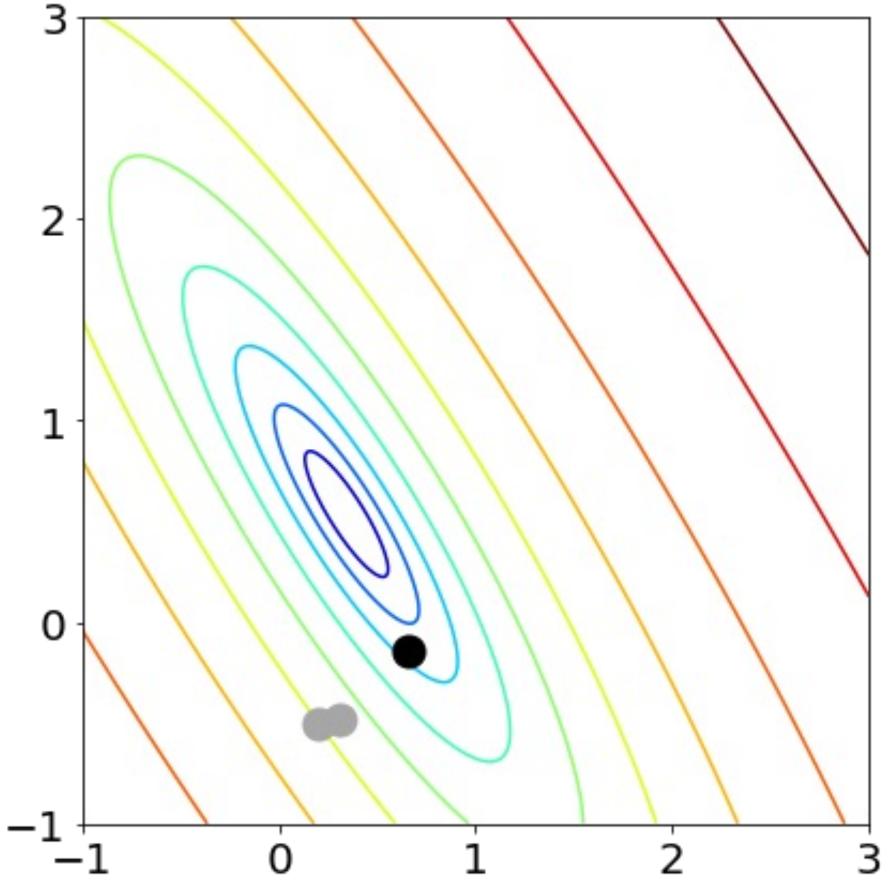


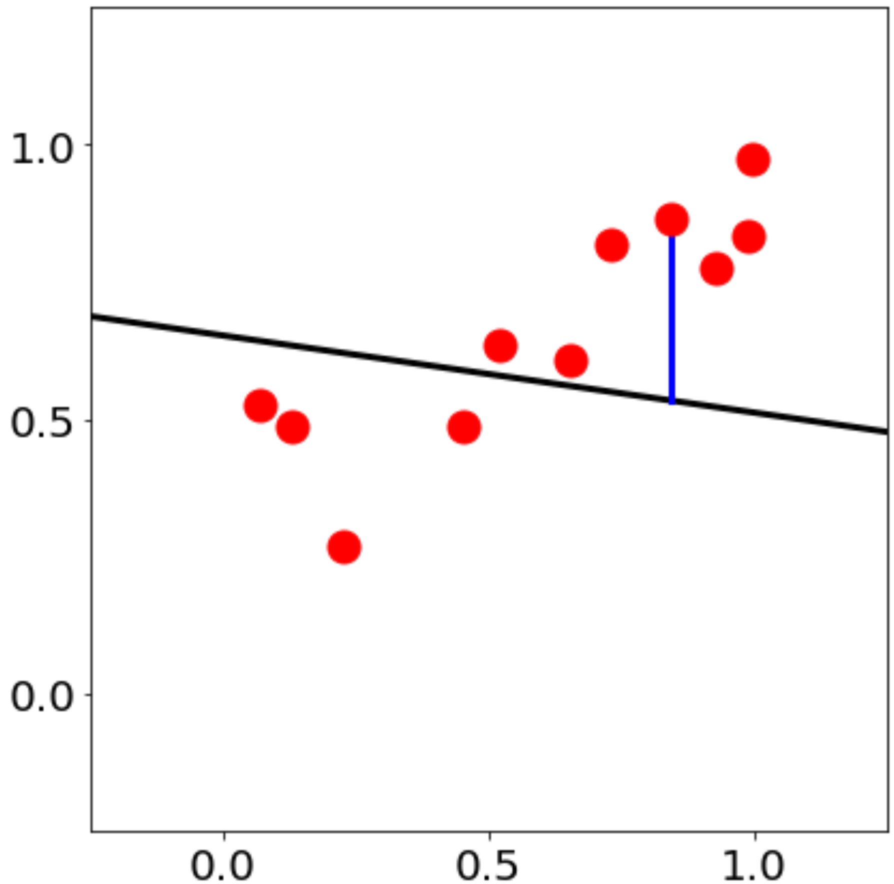


Stochastic Gradient Descent

- Initialize theta
- Select a data point & update,

```
Initialize θ
Do {
 for i=1:m
   θ ← θ - α∇θJ(i)(θ)
} while (not done)
```

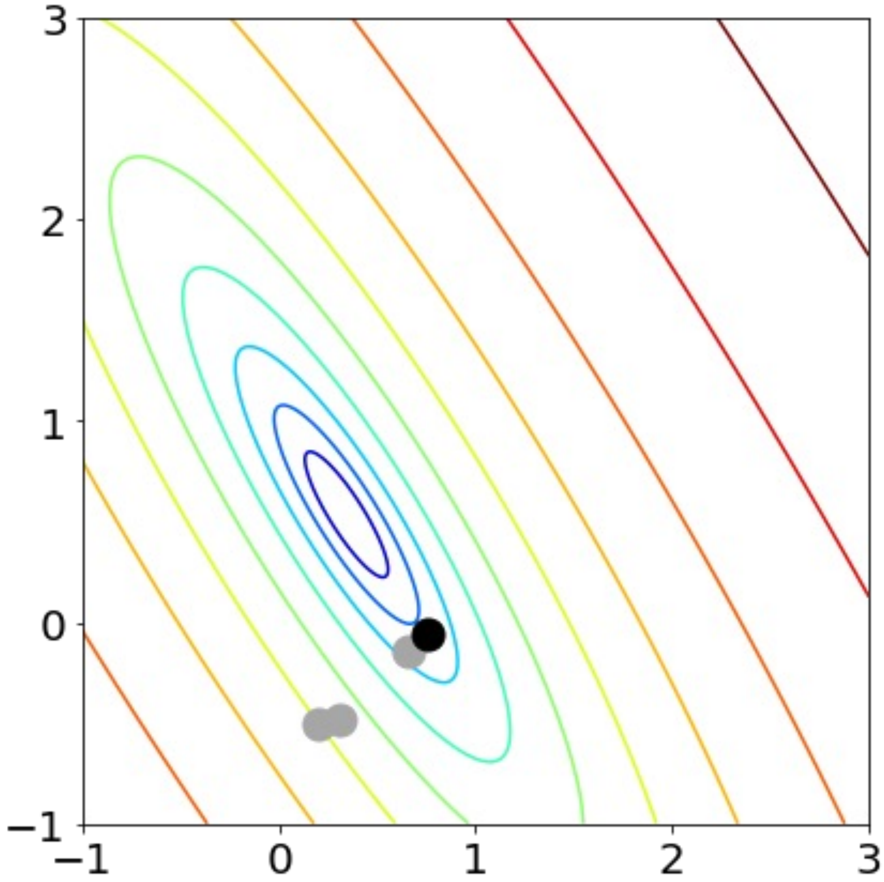


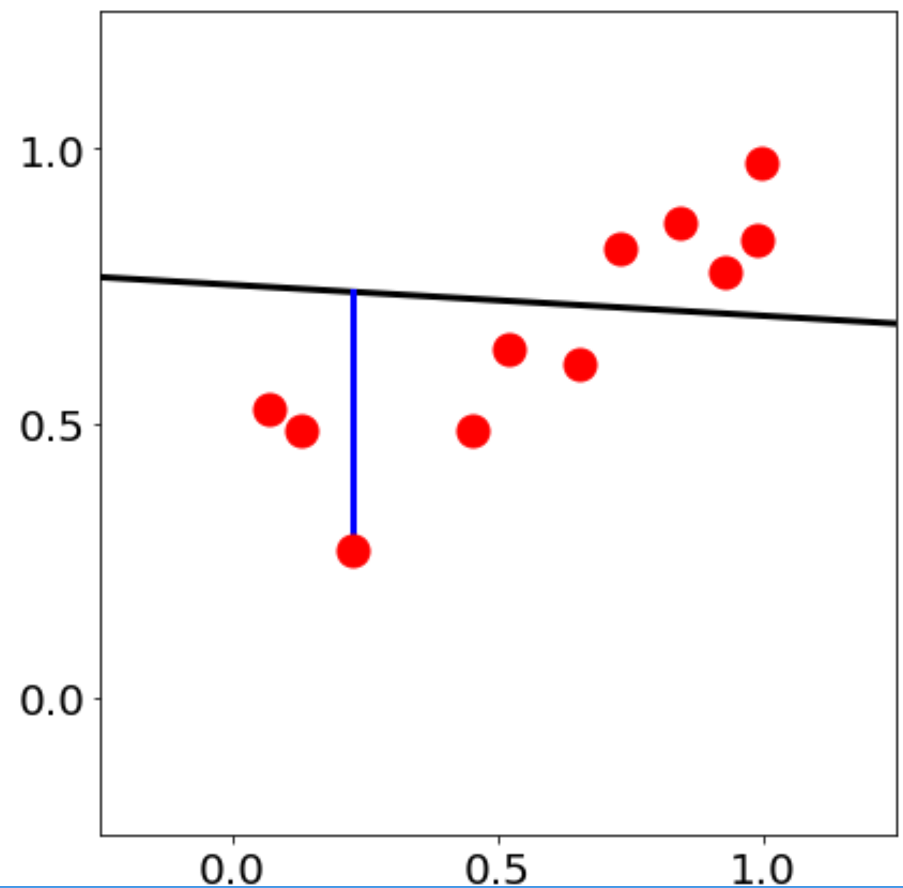


Stochastic Gradient Descent

- Initialize theta
- Select a data point & update,

```
Initialize θ
Do {
 for i=1:m
   θ ← θ - α∇θJ(i)(θ)
} while (not done)
```

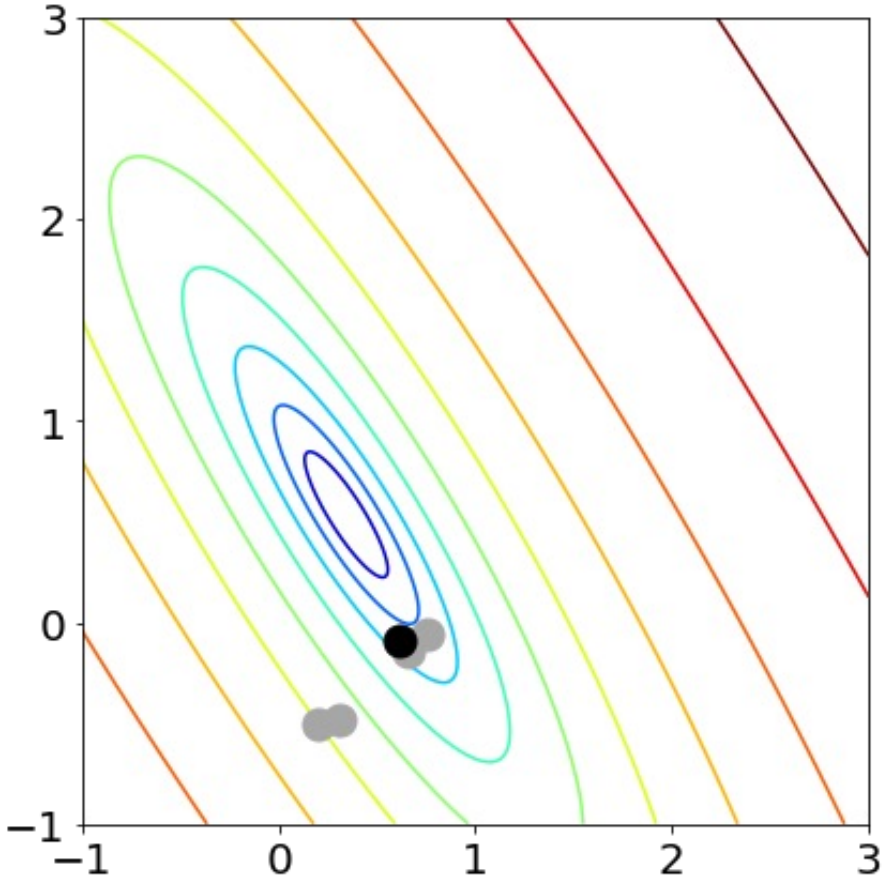


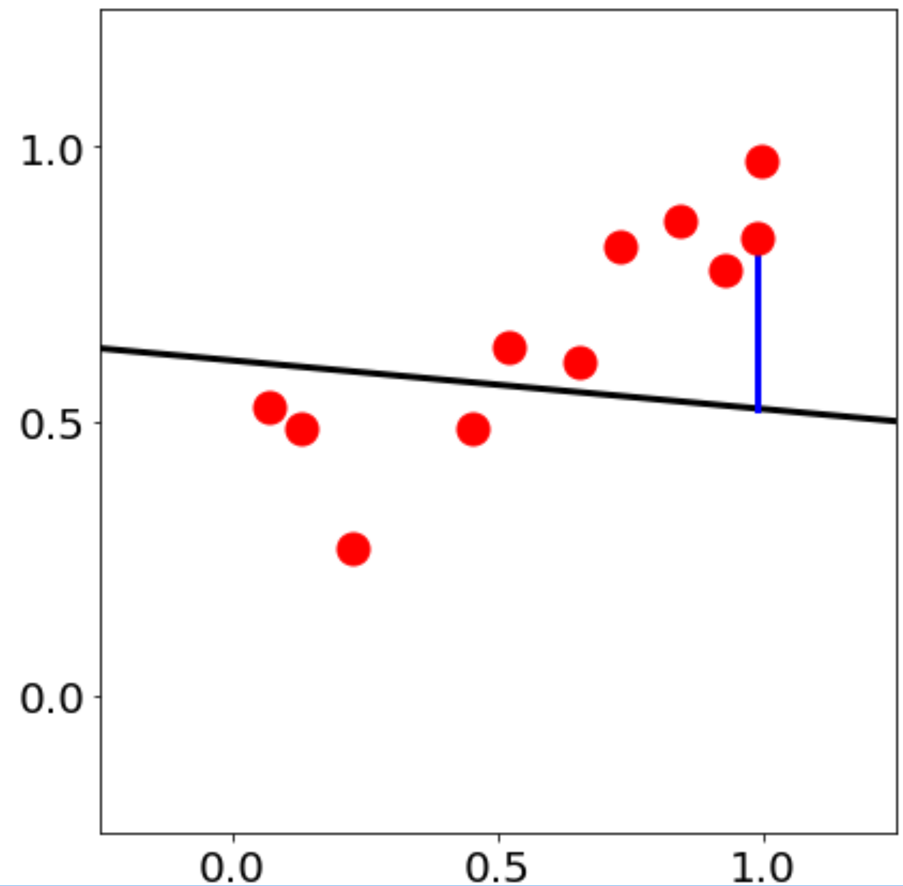


Stochastic Gradient Descent

- Initialize theta
- Select a data point & update,

```
Initialize θ
Do {
 for i=1:m
   θ ← θ - α∇θJ(i)(θ)
} while (not done)
```





Stochastic Gradient Descent

$$J^{(i)}(\theta) = \left(y^{(i)} - \theta \cdot {x^{(i)}}^T\right)^2$$

$$\nabla J^{(i)}(\theta) = 2\left(y^{(i)} - \theta \cdot {x^{(i)}}^T\right)\left[x_0^{(i)}\ ,\ x_1^{(i)}\ ,\ \ldots\right]$$

```
Initialize θ
Do {
 for i=1:m
   θ ← θ - α∇θJ(i)(θ)
} while (not done)
```

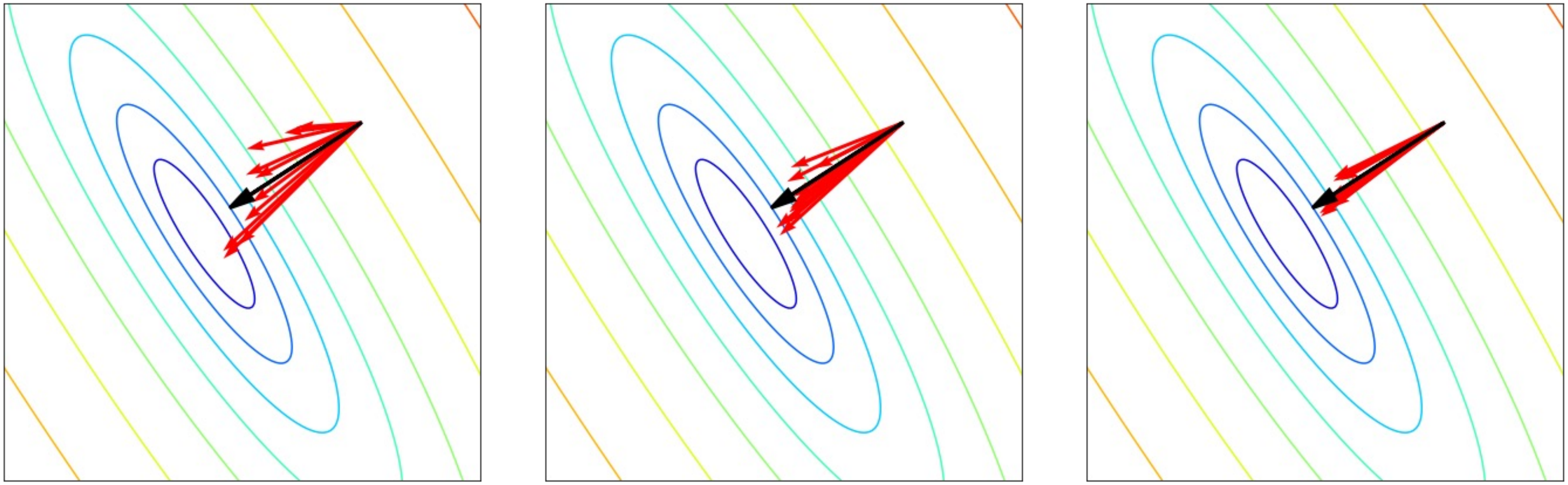
- Benefits
  - Lots of data = many more updates per pass
  - Computationally faster

- Drawbacks
  - No longer strictly “descent”
  - Stopping conditions may be harder to evaluate
    (Can use “running estimates” of J(.), etc.)

Mini-Batch Gradient Descent

- Select b data points at random & use average gradient
  - Interpolates between SGD & Batch GD
  - Fewer updates per epoch than SGD, but less noisy updates

- Example gradients (batch vs. mini-batch)



b=1 (SGD) b=3 b=5

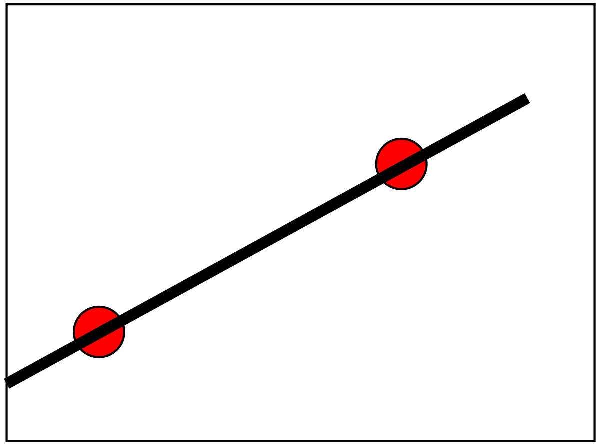
Linear Regression

- Linear Regression via Least Squares
- Gradient Descent Algorithms
- **Direct Minimization of Squared Error**
- Regression with Non-linear Features
- Bias, Variance, & Validation
- Regularized Linear Regression

MSE Minimum

- Consider a simple problem
  - One feature, two data points
  - Two unknowns: θ_0, θ_1
  - Two equations:

$$y^{(1)} = \theta_0 + \theta_1 x^{(1)}$$
$$y^{(2)} = \theta_0 + \theta_1 x^{(2)}$$



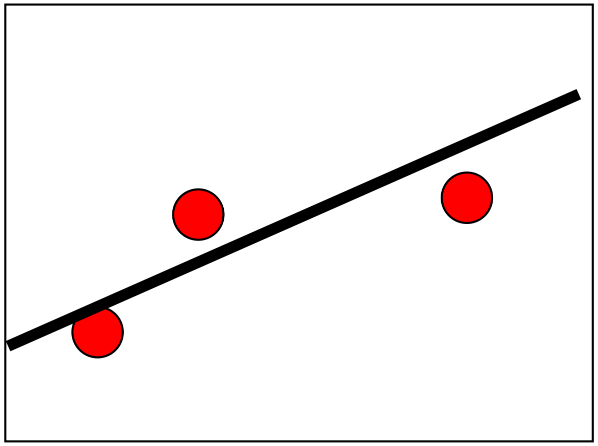
- Can solve this system directly:

$$\underline{y}^T = \underline{\theta}\,\underline{X}^T \qquad \Rightarrow \qquad \underline{\hat{\theta}} = y^T (\underline{X}^T)^{-1}$$

- However, most of the time, m > n
  - There may be no linear function that hits all the data exactly
  - Instead, solve directly for minimum of MSE function

MSE Minimum

$$\nabla J(\underline{\theta}) = -\frac{2}{m}(\underline{y}^T - \underline{\theta X}^T) \cdot \underline{X} \quad = \quad \underline{0}$$



- Simplify with some algebra:

$$\underline{y}^T \, \underline{X} - \underline{\theta X}^T \cdot \underline{X} \quad = \quad \underline{0}$$

$$\underline{y}^T \, \underline{X} = \underline{\theta X}^T \cdot \underline{X}$$

$$\underline{\theta} \quad = \quad \underline{y}^T \, \underline{X}(\underline{X}^T \, \underline{X})^{-1}$$

- $X (X^T X)^{-1}$ is called the "pseudo-inverse"

- If X^T is square and full rank, this is the inverse
- If m > n: overdetermined; gives minimum MSE fit
- $(X^T X)$ not invertible? Underdetermined (multiple sol'ns)

Python MSE

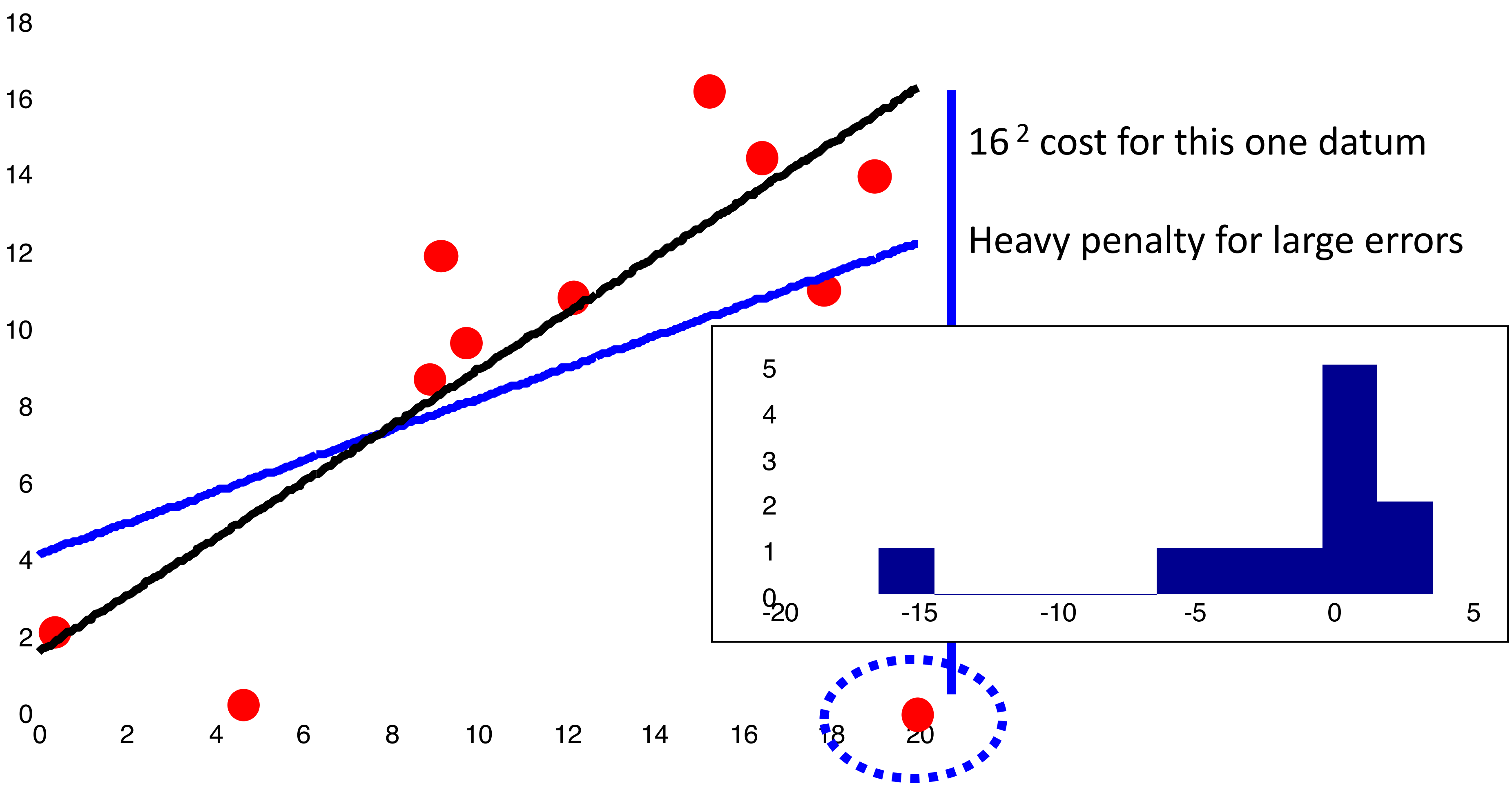
- This is easy to solve in Python / NumPy…

$$\underline{\theta} \quad = \quad \underline{y}^T \, \underline{X} (\underline{X}^T \, \underline{X})^{-1}$$

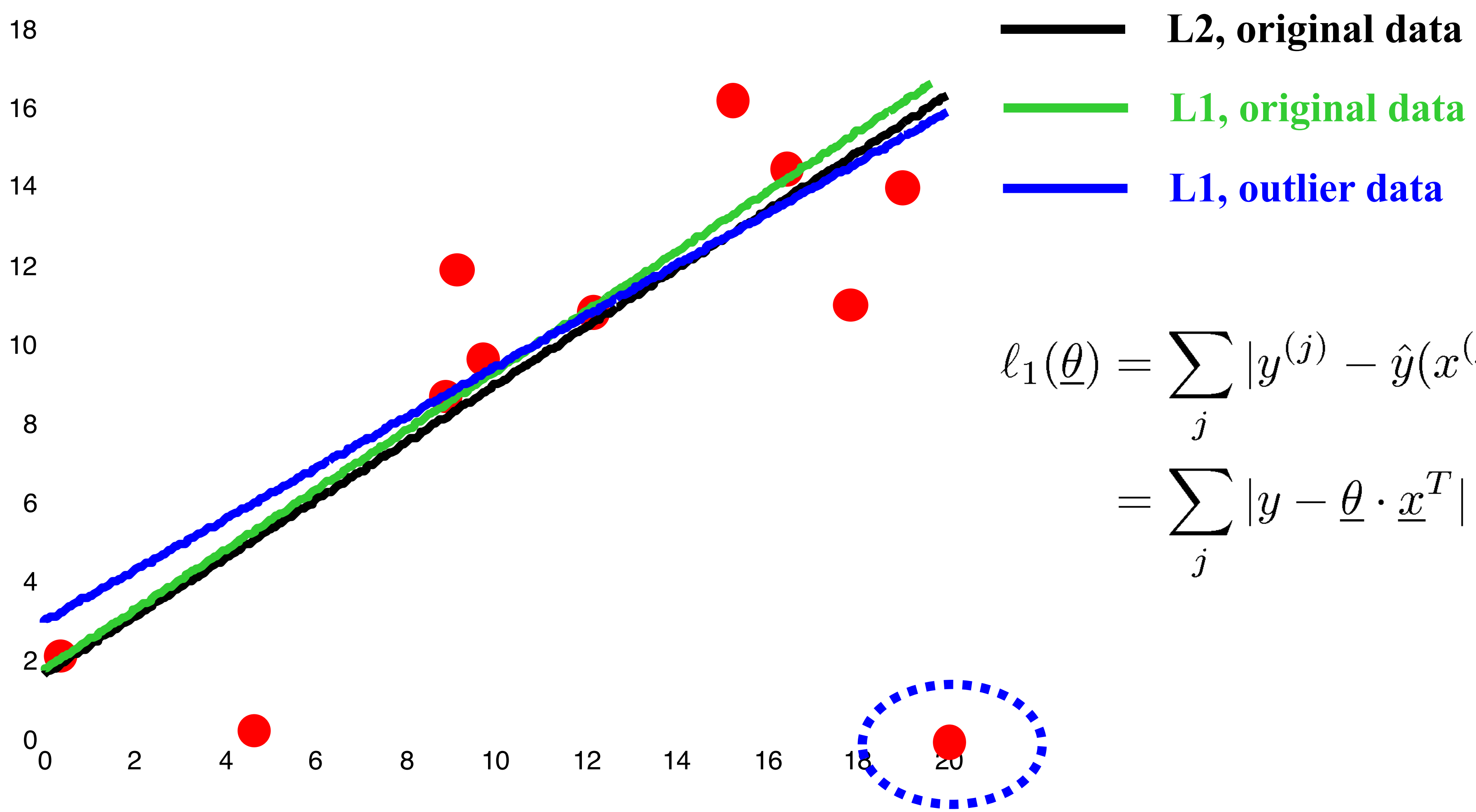
```
#  y = np.array( [[y1], … , [ym]] )
#  X = np.array( [[x1_0 … x1_n], [x2_0 … x2_n], …] )


# Solution 1: “manual”
   th = y.T @ X @ np.linalg.inv(X.T @ X)


# Solution 2: “least squares solve”
   th = np.linalg.lstsq(X, Y)
```

Effects of MSE choice

- Sensitivity to outliers



L1 error: Mean Absolute Error



$$\ell_1(\underline{\theta}) = \sum_j |y^{(j)} - \hat{y}(x^{(j)})|$$
$$= \sum_j |y - \underline{\theta} \cdot \underline{x}^T|$$

Cost functions for regression

$\ell_2 \ : \ (y - \hat{y})^2$ **(MSE)**

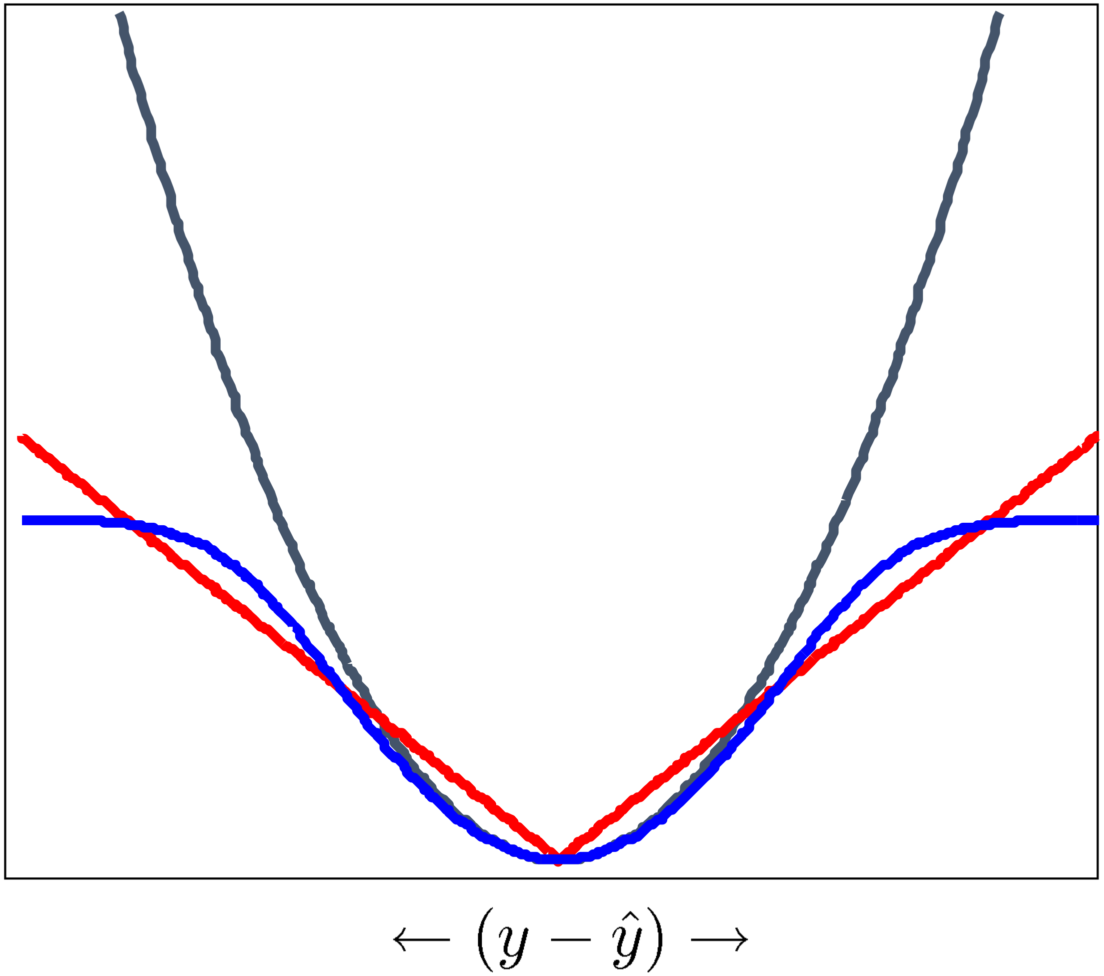
$\ell_1 \ : \ |y - \hat{y}|$ **(MAE)**

Something else entirely…

$$c - \log(\exp(-(y - \hat{y})^2) + c)$$

(???)

Arbitrary functions cannot be solved in closed form
- use gradient descent

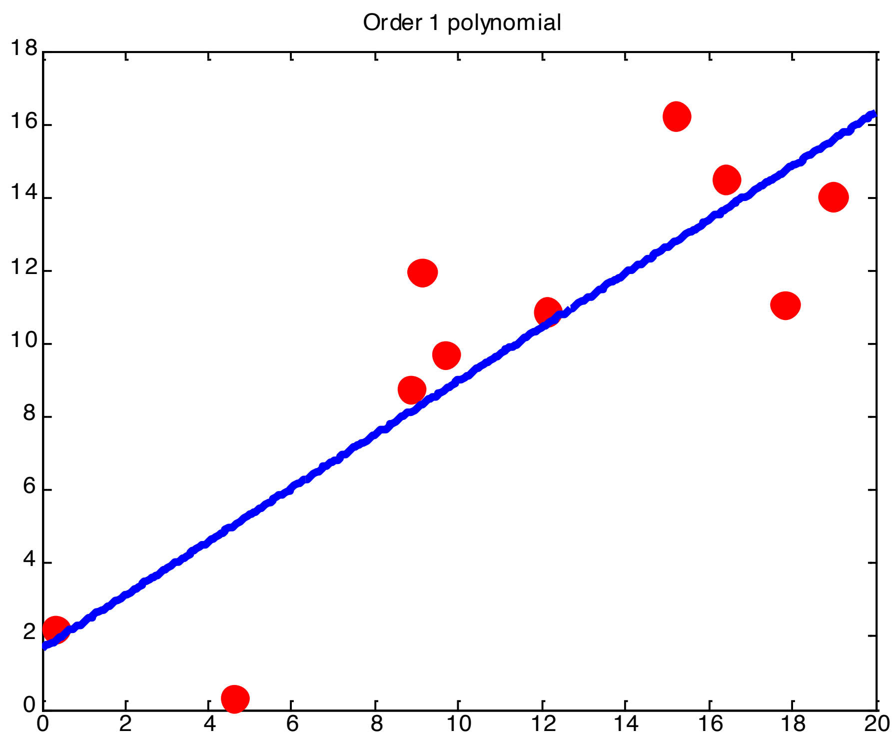


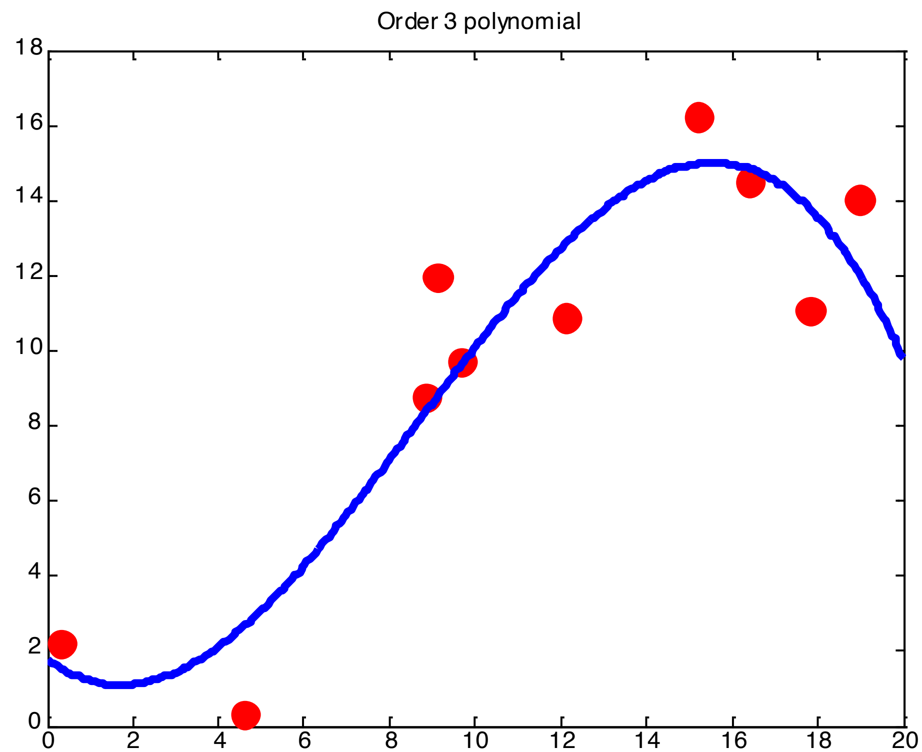
Linear Regression

- Linear Regression via Least Squares
- Gradient Descent Algorithms
- Direct Minimization of Squared Error
- **Regression with Non-linear Features**
- Bias, Variance, & Validation
- Regularized Linear Regression

Nonlinear functions

- What if our hypotheses are not lines?
  - Ex: higher-order polynomials





Nonlinear functions

- Single feature x, predict target y:

$$D = \{(x^{(j)}, y^{(j)})\}$$

$$\Downarrow$$

Add features:

$$D = \{([x^{(j)}, (x^{(j)})^2, (x^{(j)})^3], y^{(j)})\}$$

$$\hat{y}(x) = \theta_0 + \theta_1 \, x + \theta_2 \, x^2 + \theta_3 \, x^3$$

$$\Downarrow$$

$$\hat{y}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

Linear regression in new features

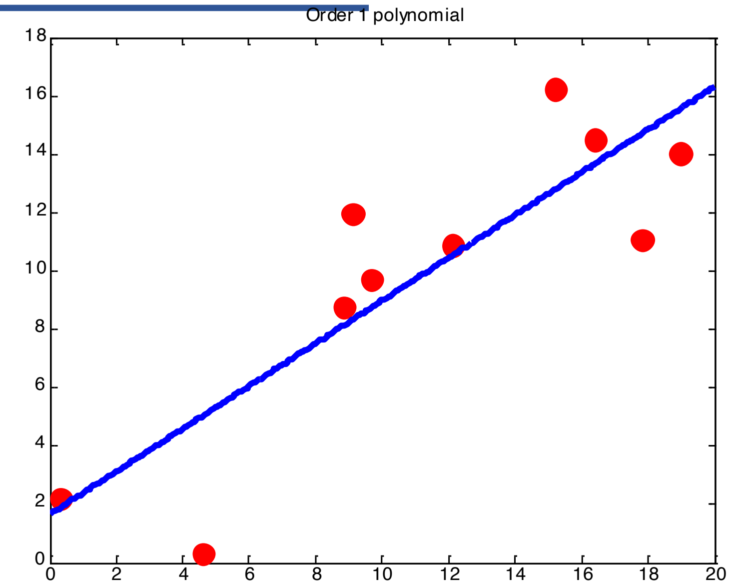
- Sometimes useful to think of "feature transform"

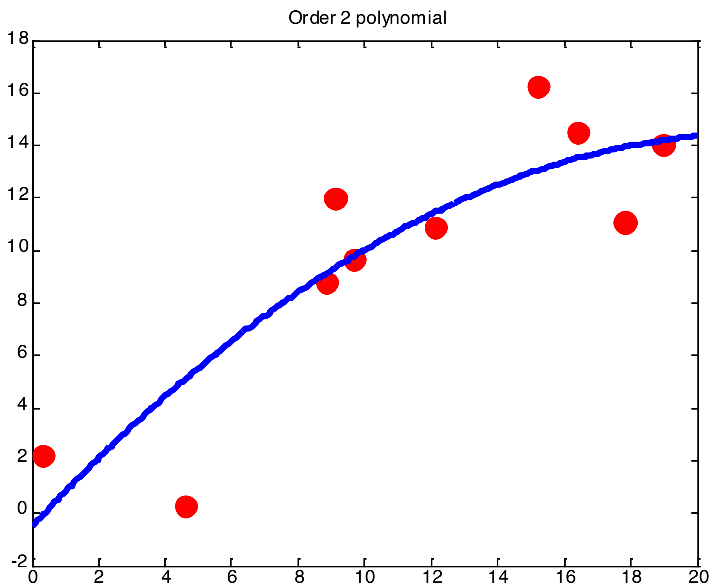
$$\Phi(x) = [1 \, , \, x \, , \, x^2 \, , \, x^3 \, , \, \ldots]$$

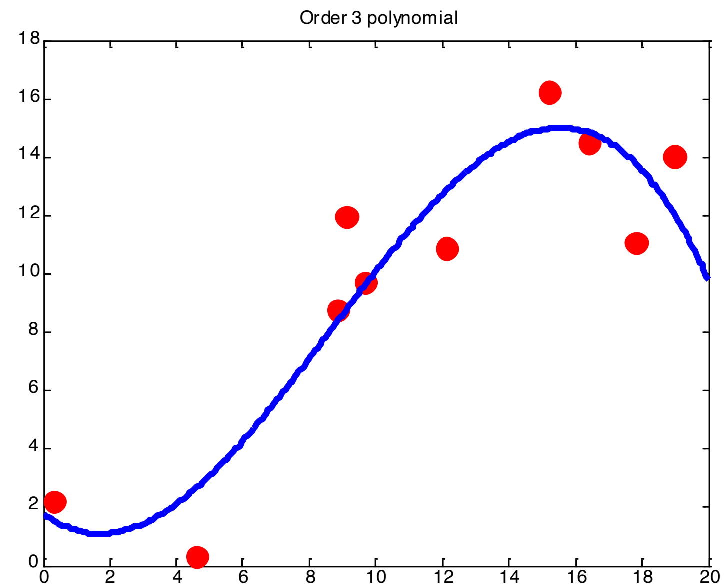
$$\hat{y}(x) = \underline{\theta} \cdot \Phi(x)$$

Higher-order polynomials

- Fit in the same way
- More "features"







Features

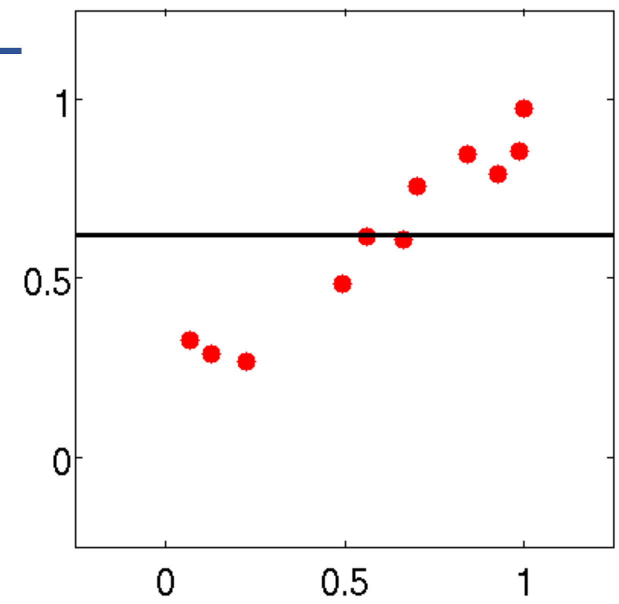
- In general, can use any features we think are useful

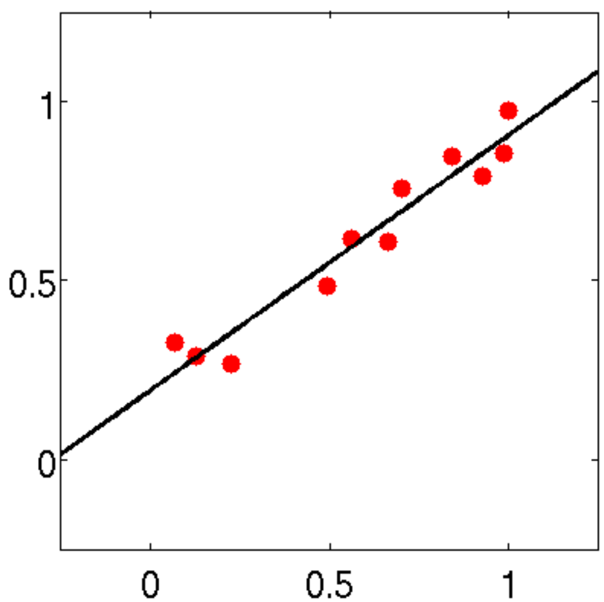
- Other information about the problem
  - Anything you can encode as fixed-length vectors of numbers
- Polynomial functions
  - Features $[1, x, x^2, x^3, \ldots]$
- Other functions
  - $1/x$, $\mathrm{sqrt}(x)$, $x_1 * x_2$, …

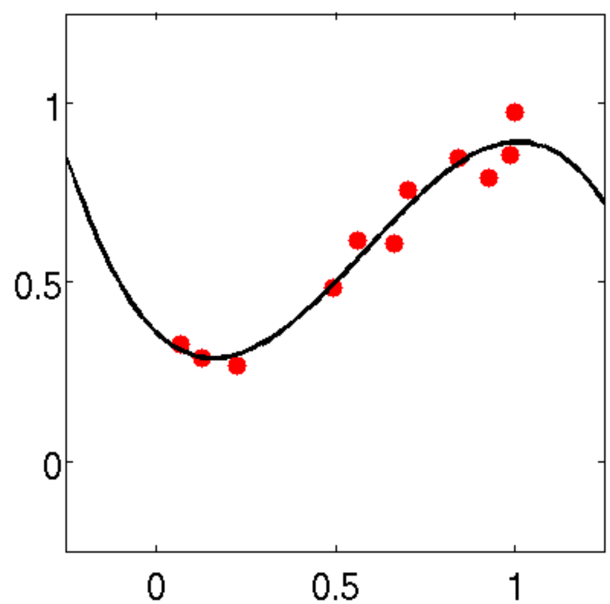
- "Linear regression" = linear in the parameters
  - Features we can make as complex as we want!

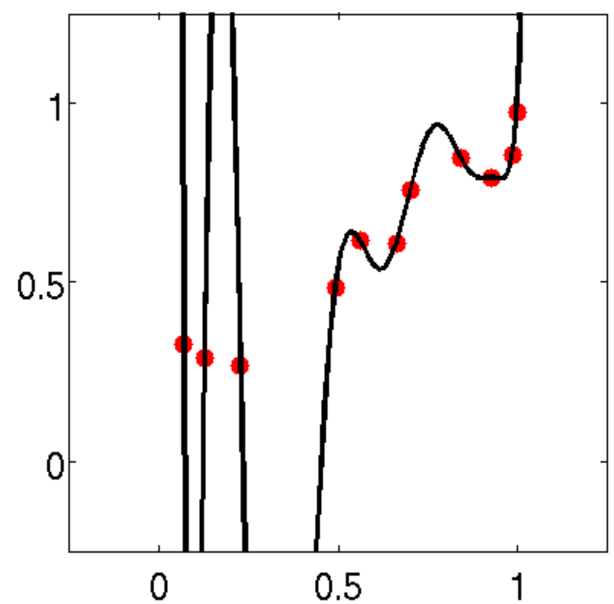
Higher-order polynomials

- Are more features better?
- "Nested" hypotheses
  - 2^{nd} order more general than 1^{st},
  - 3^{rd} order more general than 2^{nd}, …
- Fits the observed data better



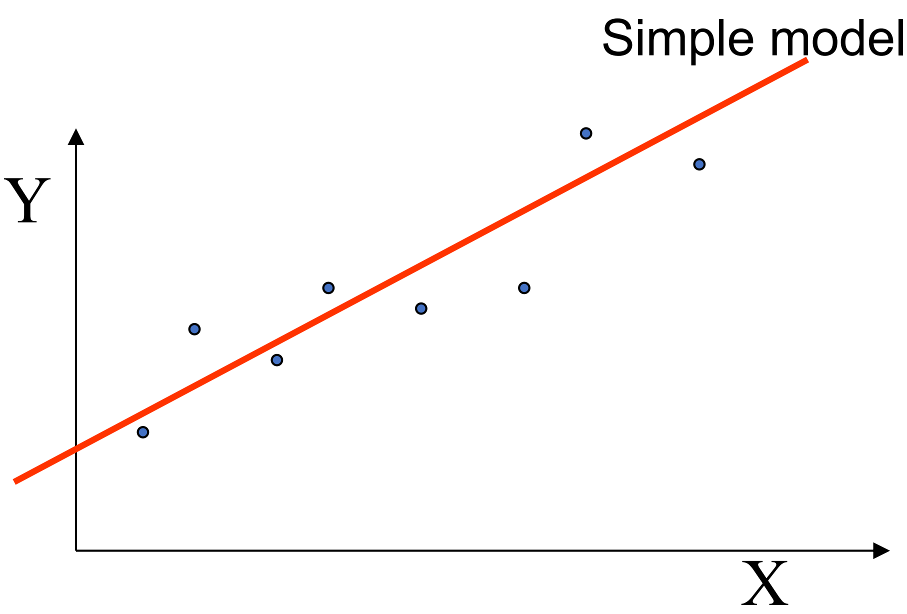


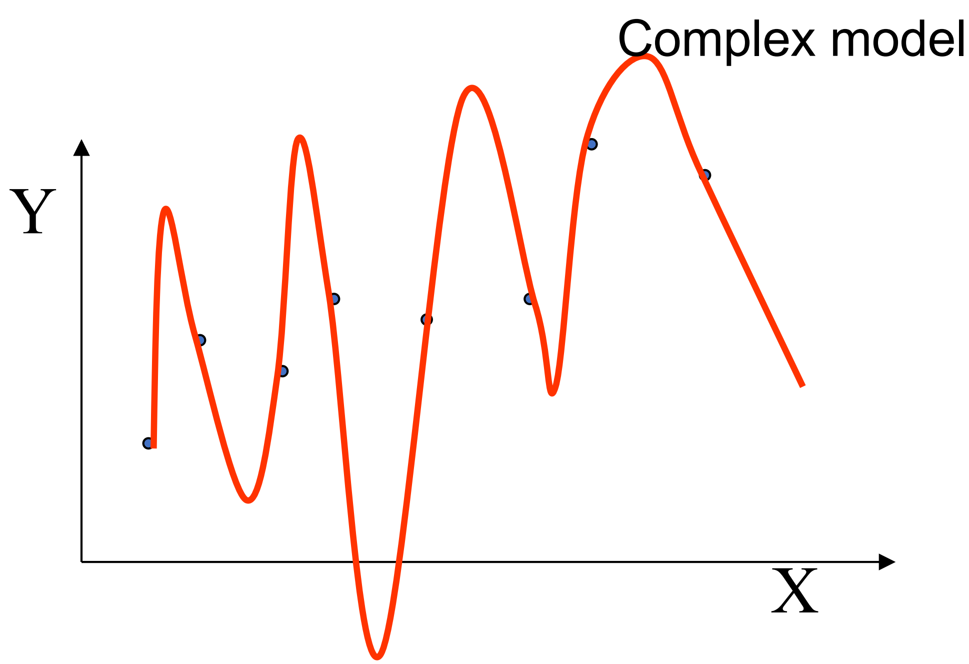




Overfitting and complexity

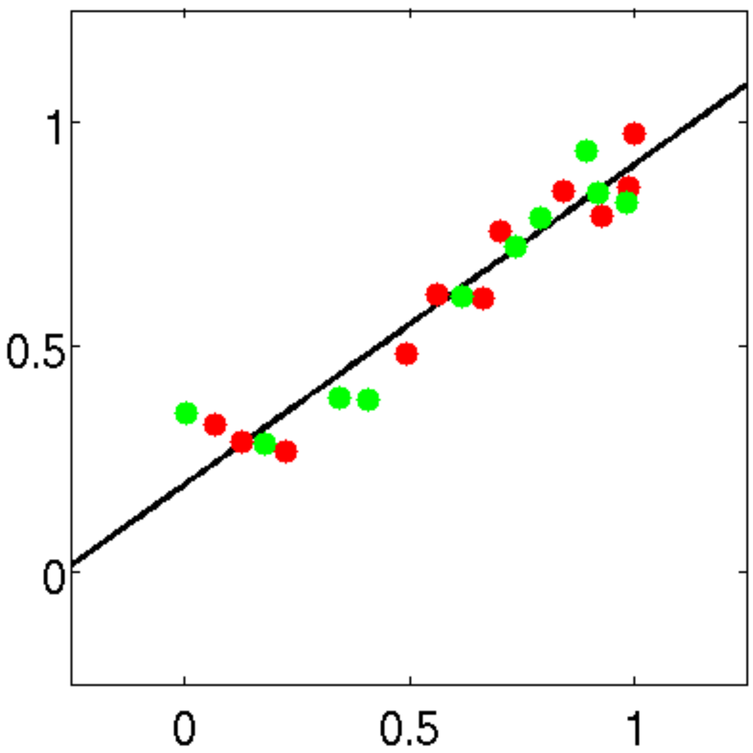
- More complex models will always fit the training data better
- But they may “overfit” the training data, learning complex relationships that are not really present

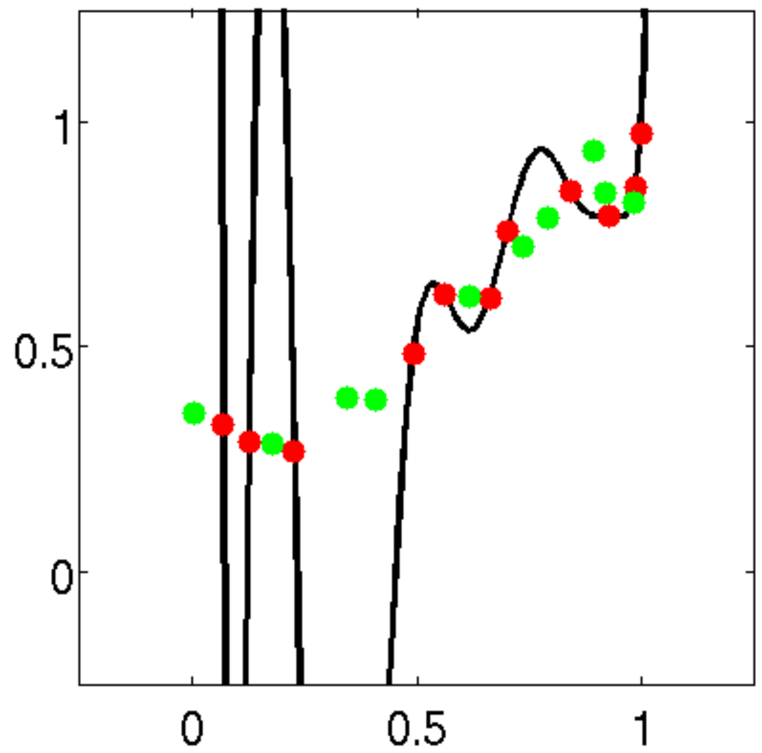




Test data

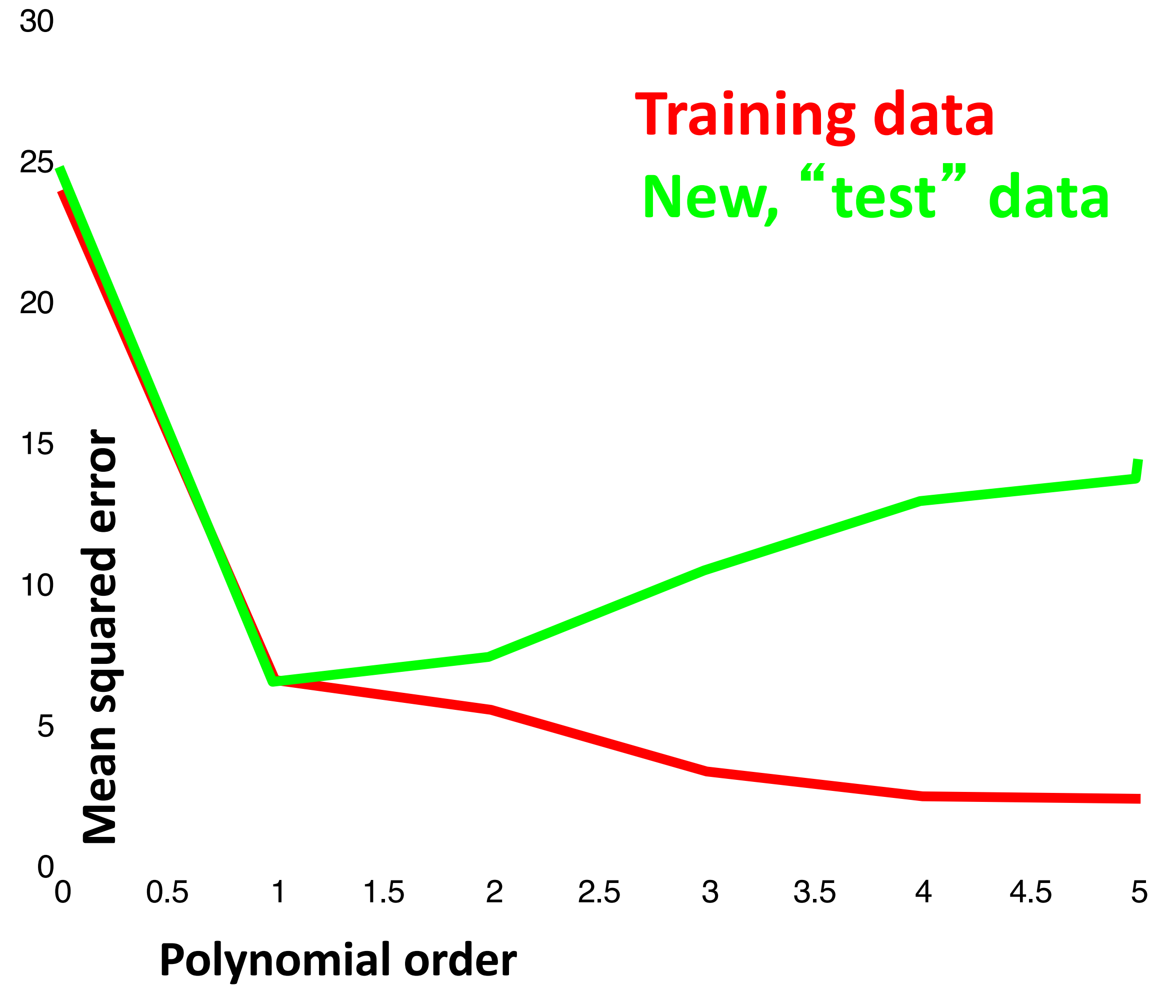
- After training the model
- Go out and get more data from the world
  - New observations (x,y)
- How well does our model perform?





Training versus test error

- Plot MSE as a function of model complexity
  - Polynomial order
- Decreases
  - More complex function fits training data better
- What about new data?
- 0^{th} to 1^{st} order
  - Error decreases
  - Underfitting
- Higher order
  - Error increases
  - Overfitting



Linear Regression

Linear Regression via Least Squares

Gradient Descent Algorithms

Direct Minimization of Squared Error

Regression with Non-linear Features

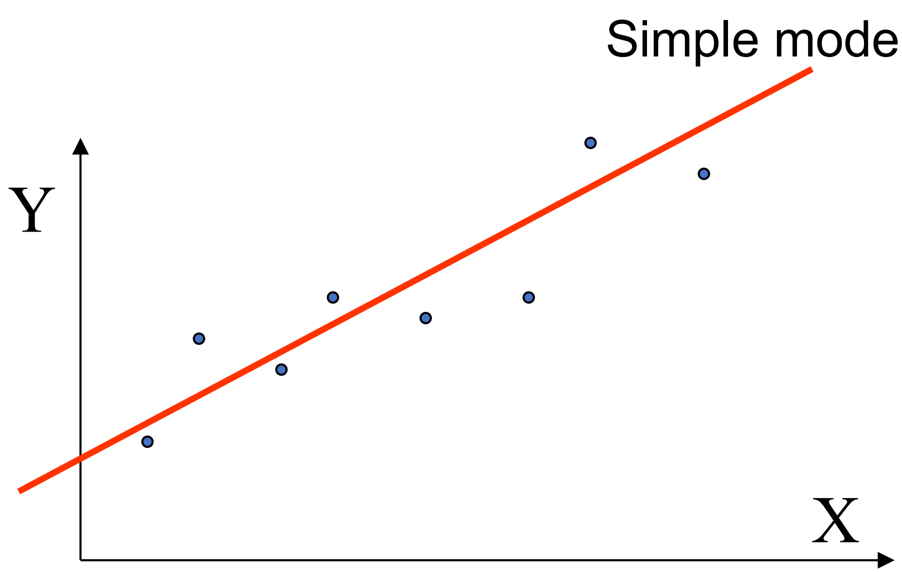
Bias, Variance, & Validation

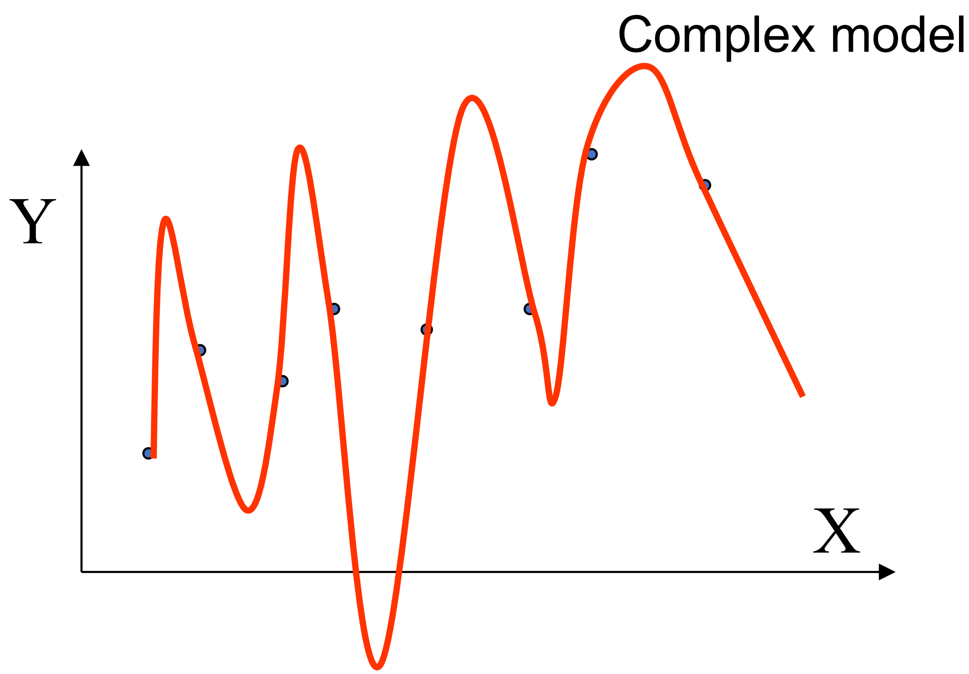
Regularized Linear Regression

Inductive bias

- The assumptions needed to predict examples we haven't seen
- Makes us "prefer" one model over another
- Polynomial functions; smooth functions; etc

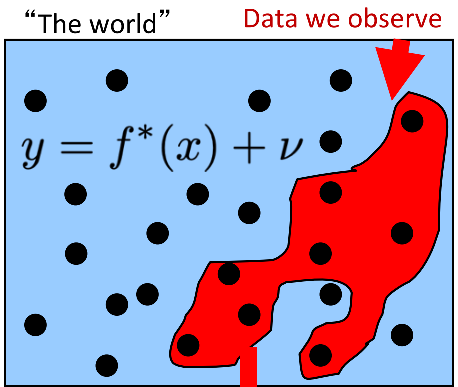
- Some bias is necessary for learning!





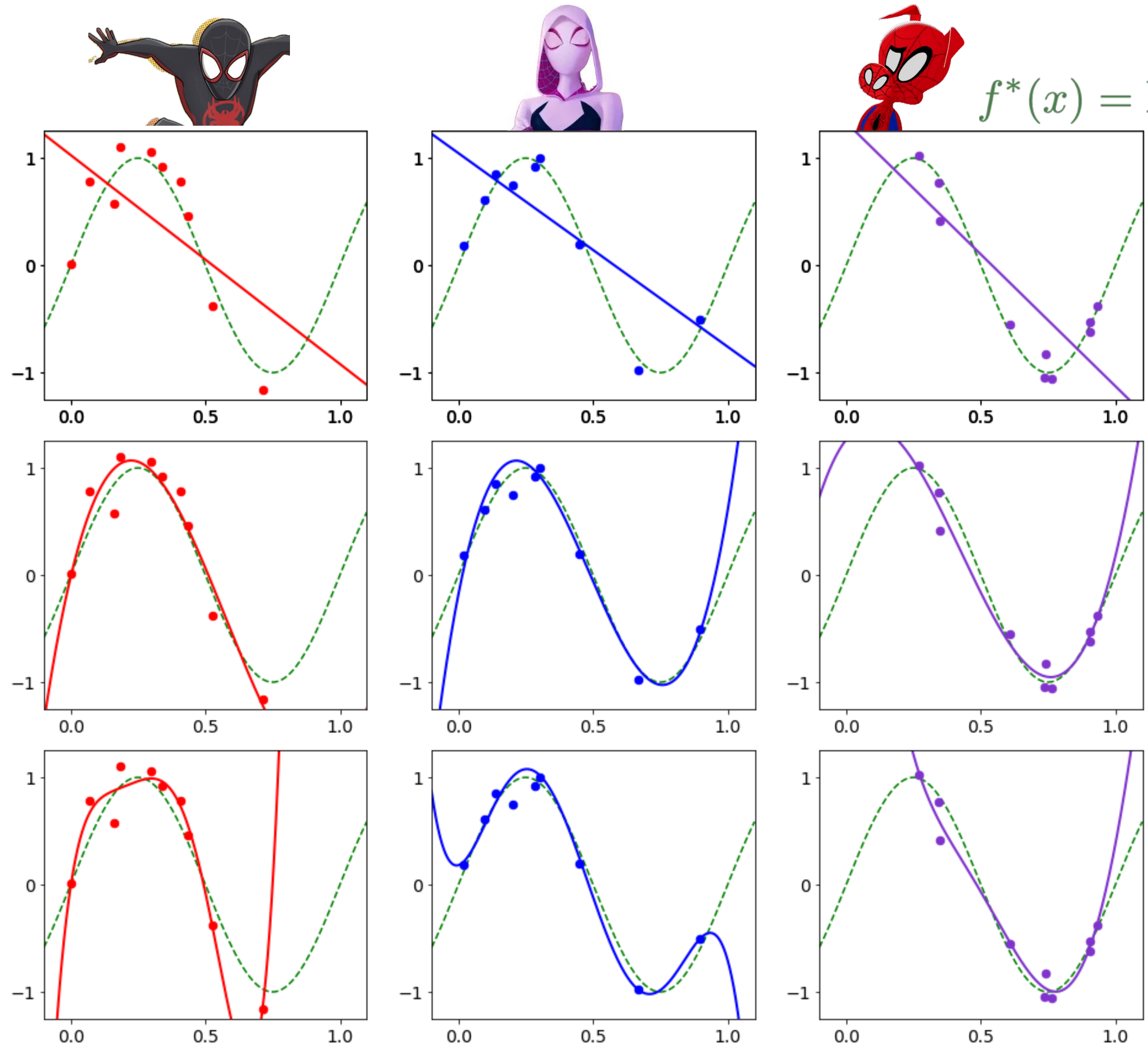
Bias & variance

We collect three different possible data sets:



$$\hat{y}(x) = \hat{\theta}_0 + \hat{\theta}_1 x$$

$$f^*(x) = \mathbb{E}[y|x]$$



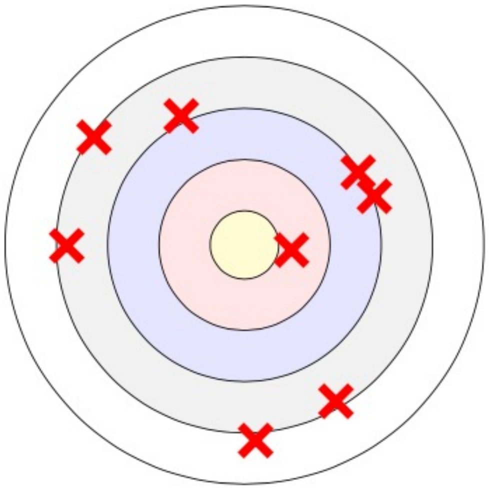
p=1

p=3

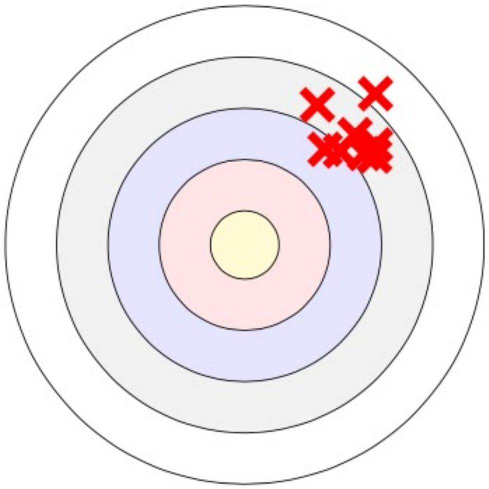
p=5

Bias & variance

- Two different types of "errors"



Low bias, high variance



High bias, low variance

- Same MSE (average distance to the bullseye) in both!

$$\frac{1}{m}\sum_i \|x^{(i)} - x^*\|^2 = \frac{1}{m}\sum_i \|\bar{x} - x^*\|^2 + \frac{1}{m}\sum_i \|x^{(i)} - \bar{x}\|^2$$

MSE

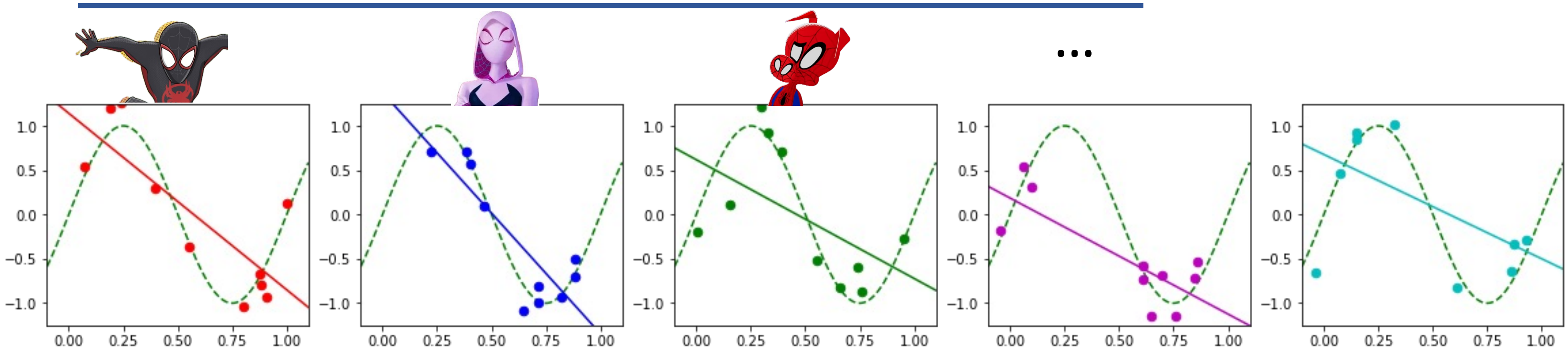
(Bias)^2

Systematic error, "on average"

Variance

Variation from run to run

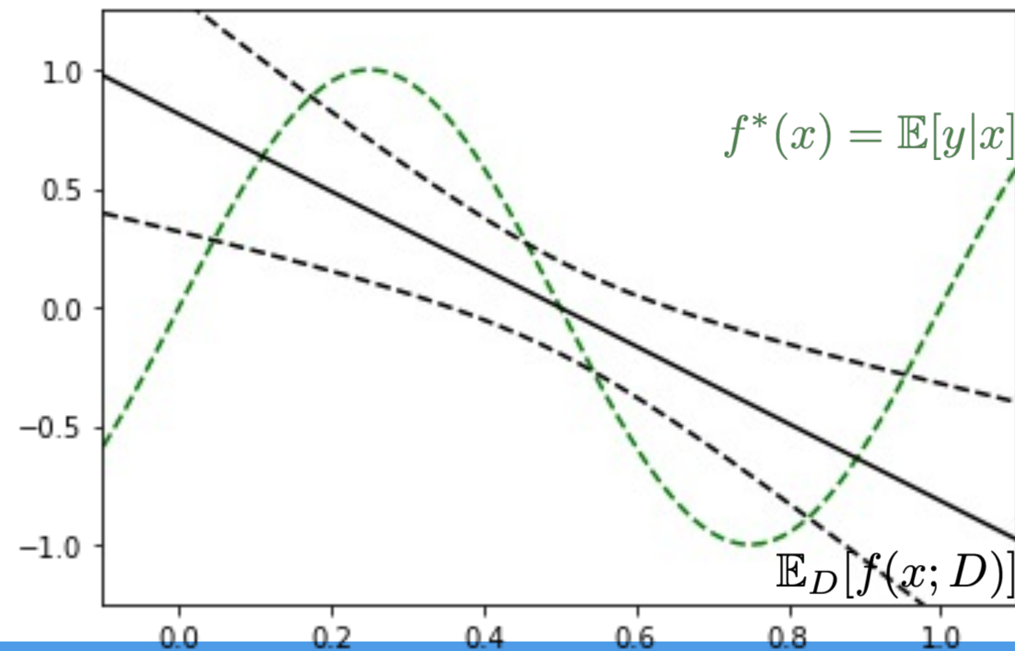
Bias & variance



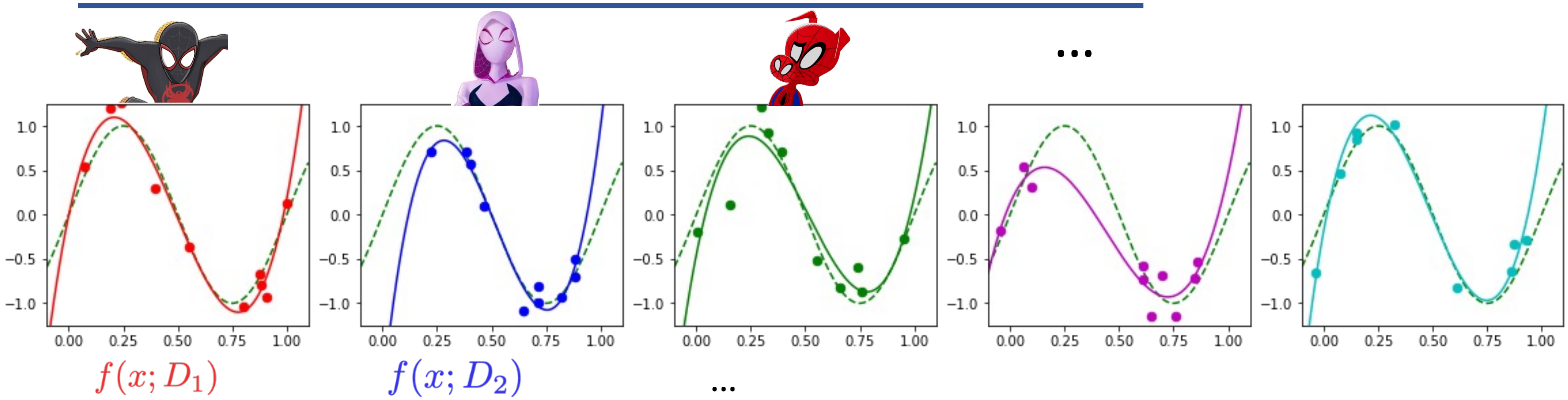
$f(x; D_1)$ $f(x; D_2)$ …

Degree = 1
M = 10 data points

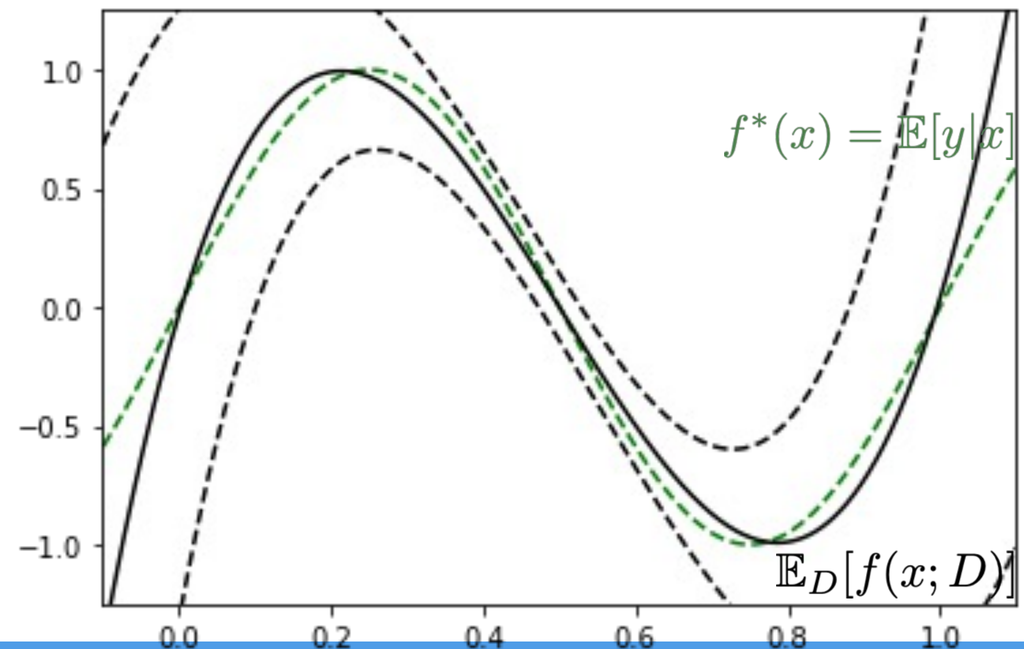
All models are similar to each other
Models are not very close to f*



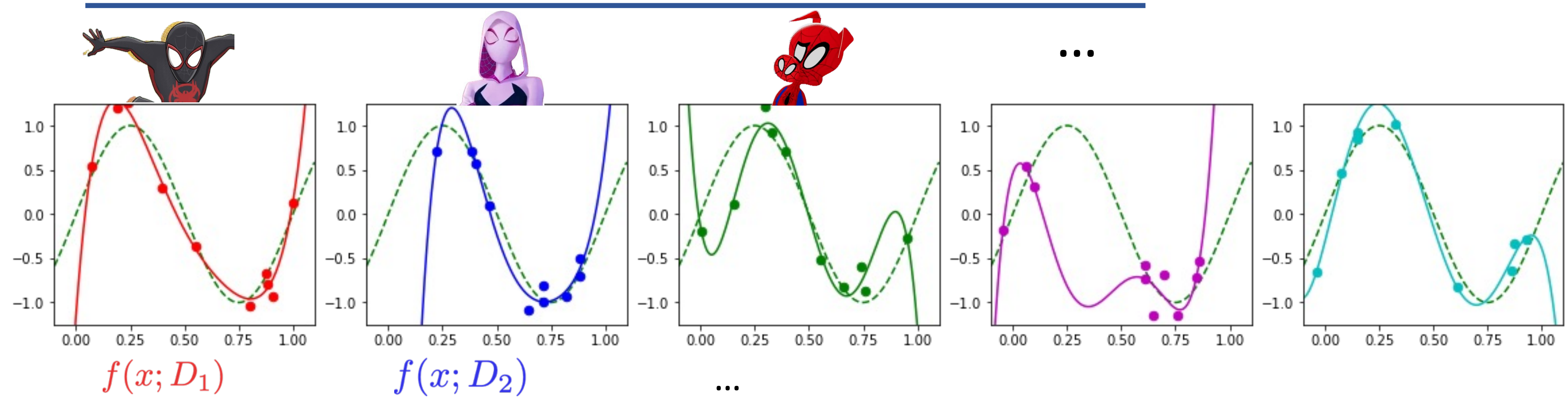
Bias & variance



Degree = 3
m = 10 data points

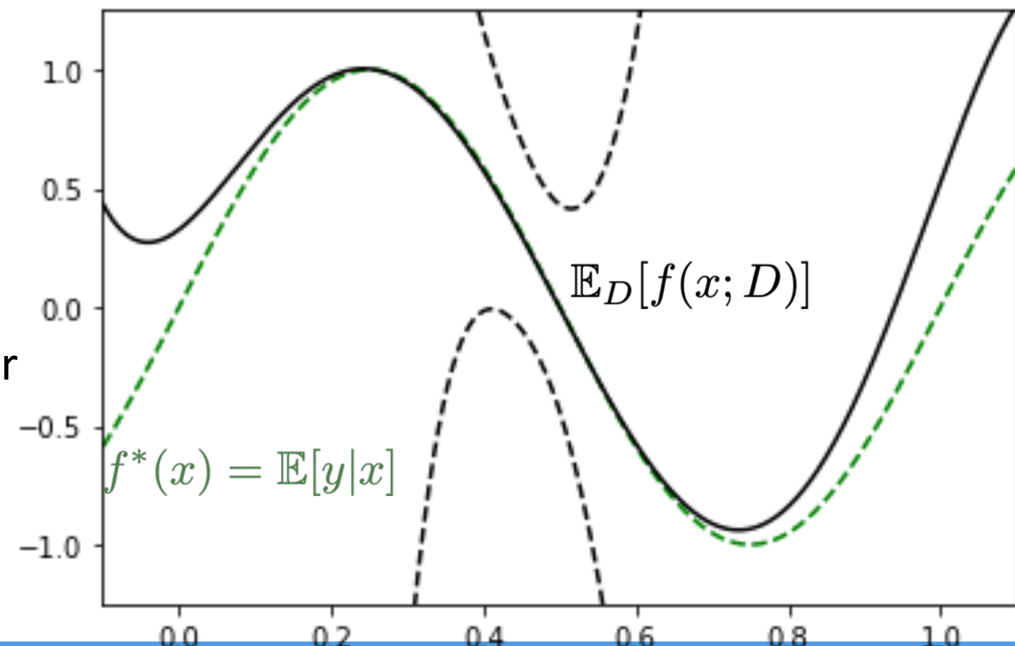


Bias & variance



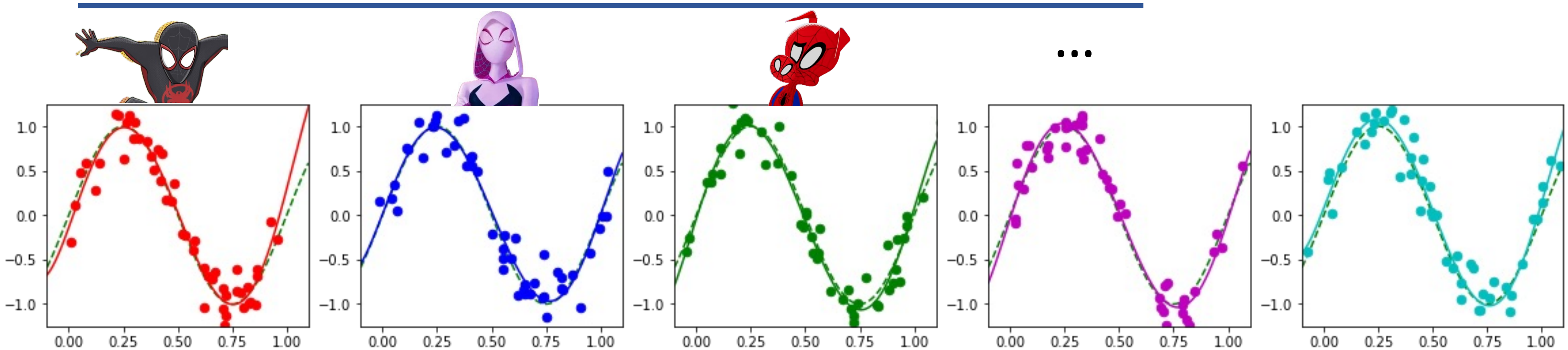
Degree = 6
m = 10 data points

Models are, **on average**, close to f*
Models are very different from each other



Bias & variance

…

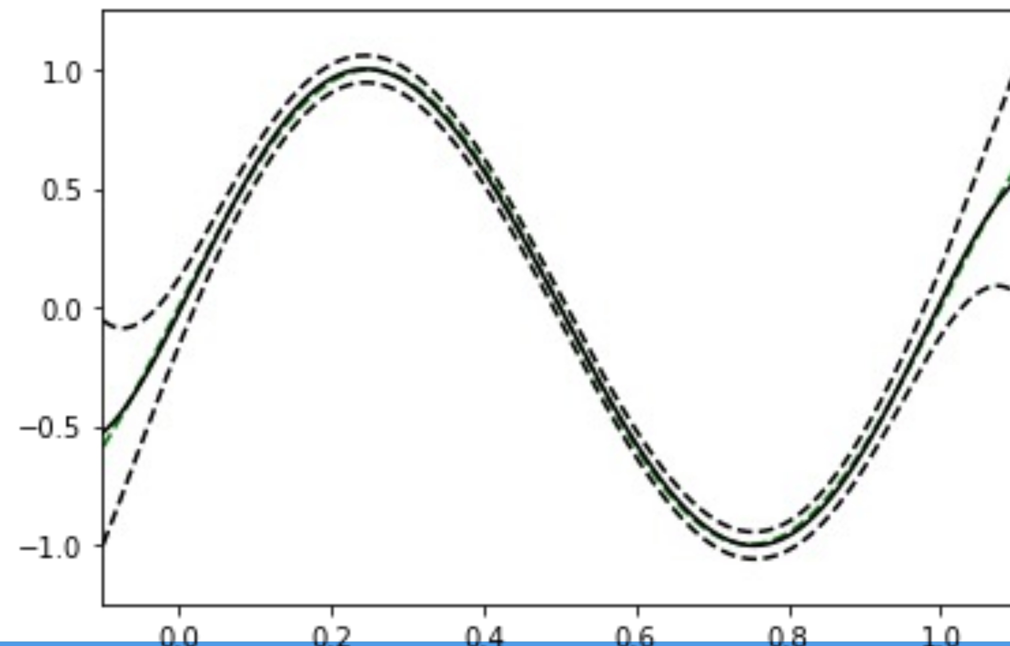


$f(x; D_1)$ $f(x; D_2)$ …

Degree = 6
m = 50 data points

Bias & variance are functions of the learner, training procedure, and training data size

Here: given more data per model, the models are far more similar (lower variance), while still having low bias



Detecting overfitting

- Overfitting effect
  - Do better on training data than on future data
  - Need to choose the “right” complexity

- One solution: “Hold-out” data
- Separate our data into two sets
  - Training
  - Test
- Learn only on training data
- Use test data to estimate generalization quality
  - Model selection

- All good competitions use this formulation
  - Often multiple splits: one by judges, then another by you

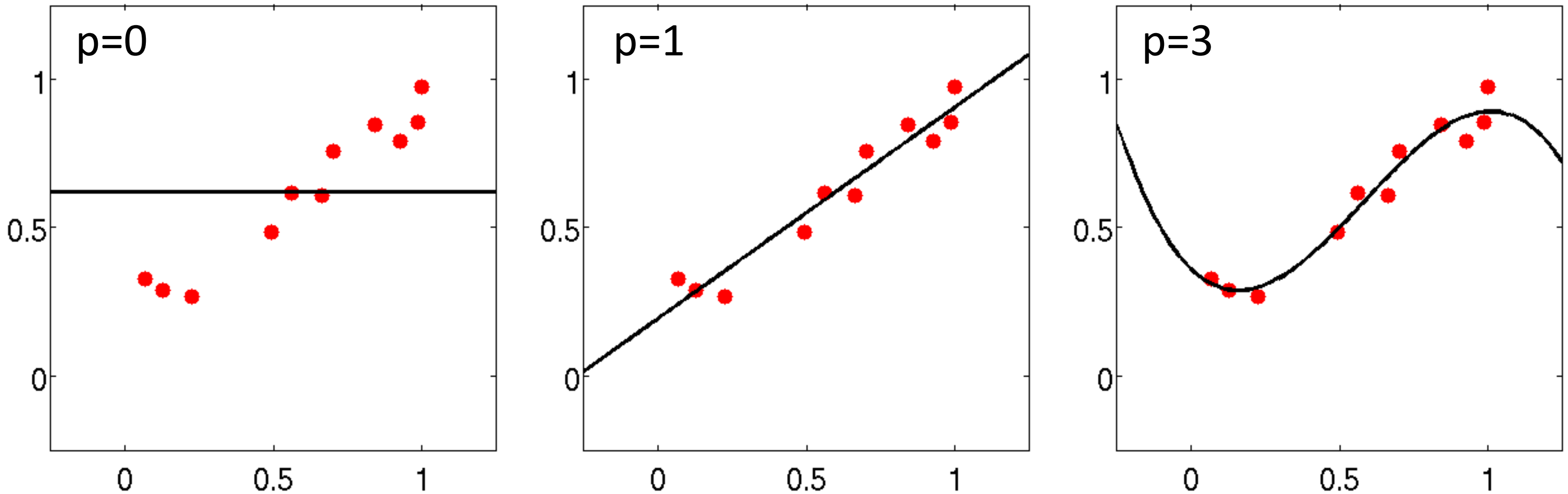
Model selection

- Which of these models fits the data best?
  - p=0 (constant); p=1 (linear); p=3 (cubic); …
- Or, should we use KNN? Other methods?

- Model selection problem
  - Can't use training data to decide (esp. if models are nested!)

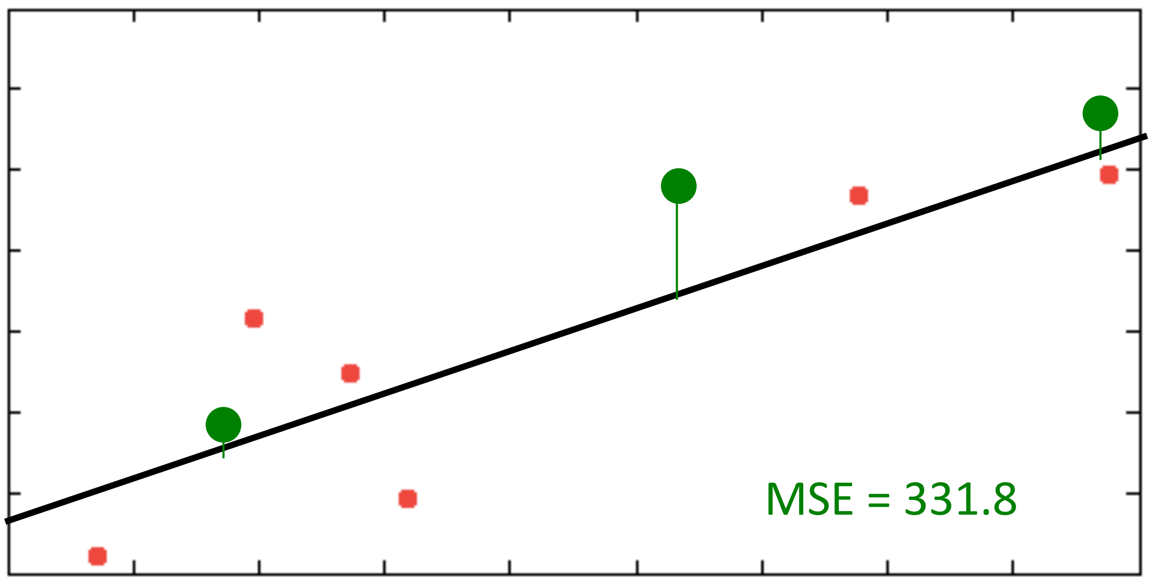
- Want to estimate $\mathbb{E}_{(x,y)}[J(y, \hat{y}(x\,;\,D))]$

J = loss function (MSE)
D = training data set



Hold-out method

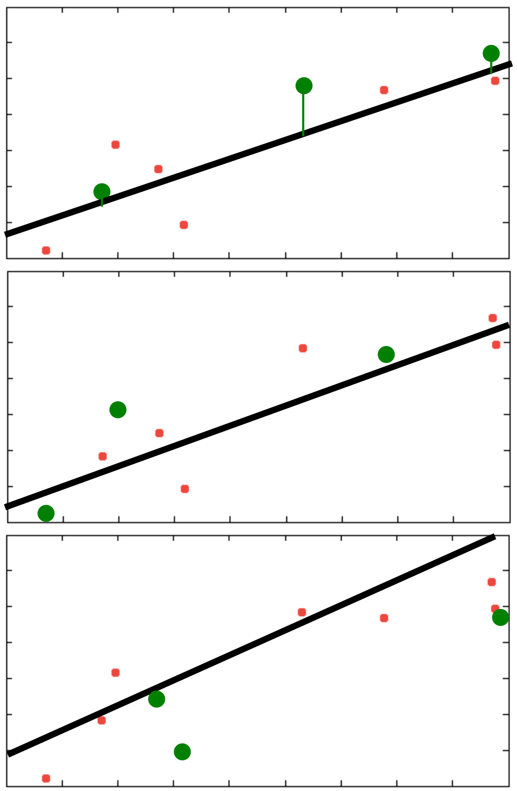
- Validation data
  - “Hold out” some data for evaluation (e.g., 70/30 split)
  - Train only on the remainder
- Some problems, if we have few data:
  - Few data in hold-out: noisy estimate of the error
  - More hold-out data leaves less for training!



	$x^{(i)}$	$y^{(i)}$
Training data	88	79
	32	-2
	27	30
	68	73
	7	-16
	20	43
Validation data	53	77
	17	16
	87	94

Cross-validation method

- K-fold cross-validation
  - Divide data into K disjoint sets
  - Hold out one set (= M / K data) for evaluation
  - Train on the others (= M*(K-1) / K data)



Split 1:
MSE = 331.8

Split 2:
MSE = 361.2

Split 3:
MSE = 669.8

3-Fold X-Val MSE
= 464.1

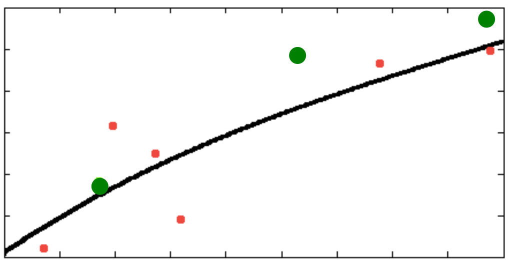
Training data

Validation data

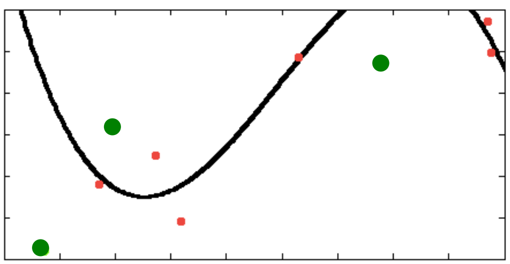
$x^{(i)}$	$y^{(i)}$
88	79
32	-2
27	30
68	73
7	-16
20	43
53	77
17	16
87	94

Cross-validation method

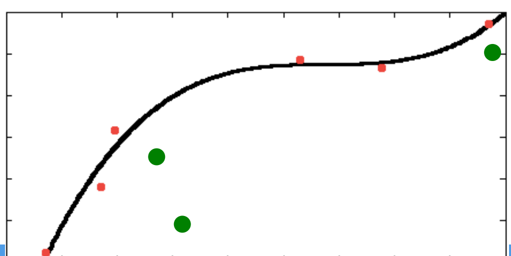
- K-fold cross-validation
  - Divide data into K disjoint sets
  - Hold out one set (= M / K data) for evaluation
  - Train on the others (= M*(K-1) / K data)



Split 1:
MSE = 280.5



Split 2:
MSE = 3081.3



Split 3:
MSE = 1640.1

3-Fold X-Val MSE
= 1667.3

Training data

Validation data

$x^{(i)}$	$y^{(i)}$
88	79
32	-2
27	30
68	73
7	-16
20	43
53	77
17	16
87	94

Cross-validation

- Advantages:
  - Lets us use more (M) validation data
    (= less noisy estimate of test performance)

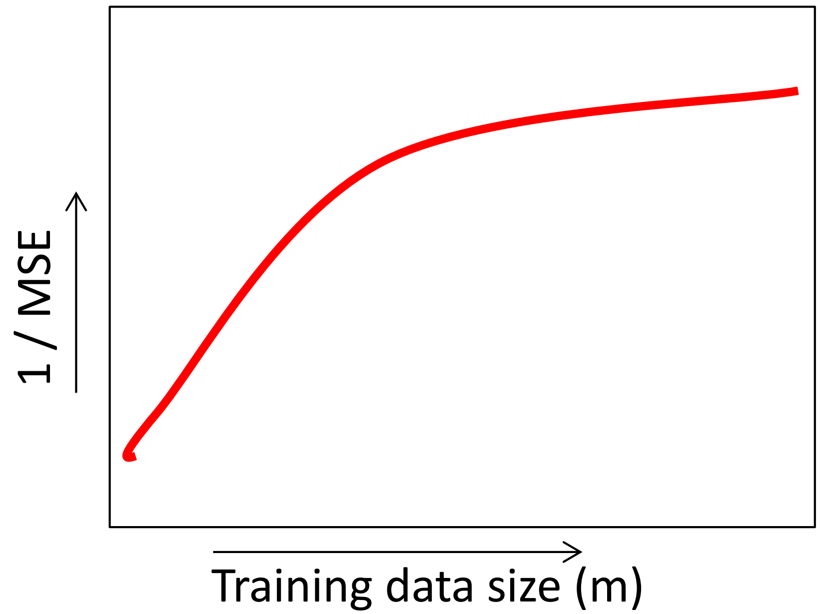
- Disadvantages:
  - More work
    - Trains K models instead of just one
  - Doesn't evaluate any *particular* predictor
    - Evaluates K different models & averages
    - Scores *hyperparameters / procedure*, not an actual, specific predictor!

- Also: still estimating error for M' < M data…

Learning curves

- Plot performance as a function of training size
  - Assess impact of fewer data on performance
    Ex: MSE0 - MSE (regression)
    or 1-Err (classification)

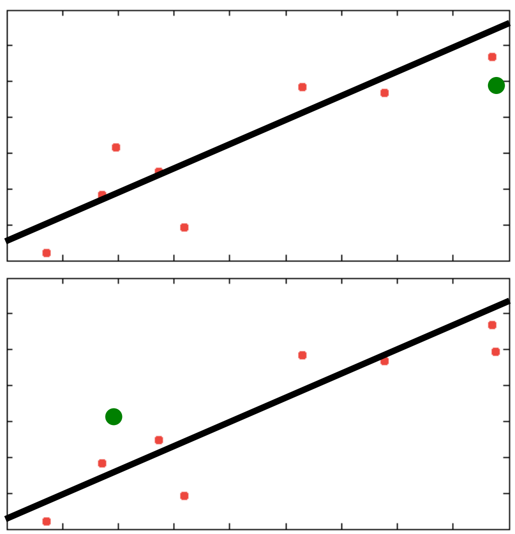
- Few data
  - More data significantly improve performance
- "Enough" data
  - Performance saturates



- If slope is high, decreasing *m* (for validation / cross-validation) might have a big impact…

Leave-one-out cross-validation

- When K=M (# of data), we get
  - Train on all data except one
  - Evaluate on the left-out data
  - Repeat M times (each data point held out once) and average



MSE = …

MSE = …

⋮

LOO X-Val MSE = …

Training data

Validation data

$x^{(i)}$	$y^{(i)}$
88	79
32	-2
27	30
68	73
7	-16
20	43
53	77
17	16
87	94

Cross-validation Issues

- Need to balance:
  - Computational burden (multiple trainings)
  - Accuracy of estimated performance / error
- Single hold-out set:
  - Estimates performance with M' < M data (important? learning curve?)
  - Need enough data to trust performance estimate
  - Estimates performance of a particular, trained learner
- K-fold cross-validation
  - K times as much work, computationally
  - Better estimates, still of performance with M' < M data
- Leave-one-out cross-validation
  - M times as much work, computationally
  - M' = M-1, but overall error estimate may have high variance

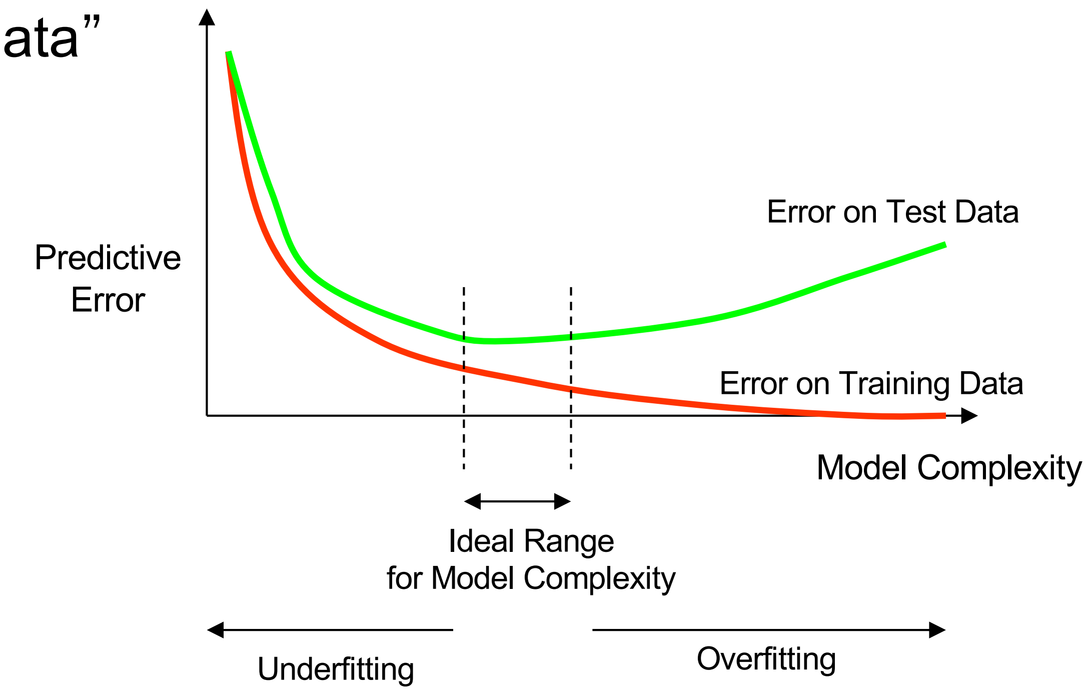
Linear Regression

- Linear Regression via Least Squares
- Gradient Descent Algorithms
- Direct Minimization of Squared Error
- Regression with Non-linear Features
- Bias, Variance, & Validation
- **Regularized Linear Regression**

What to do about under/overfitting?

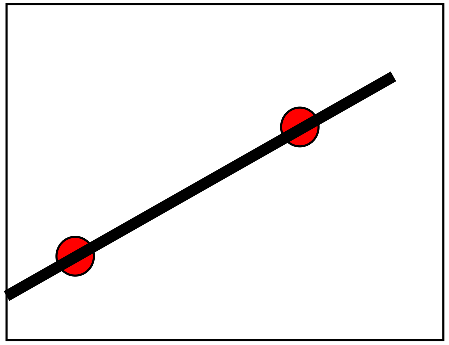
- Ways to increase complexity?
  - Add features, parameters
  - We'll see more…

- Ways to decrease complexity?
  - Remove features ("feature selection")
  - "Fail to fully memorize data"
    - Partial training
    - Regularization

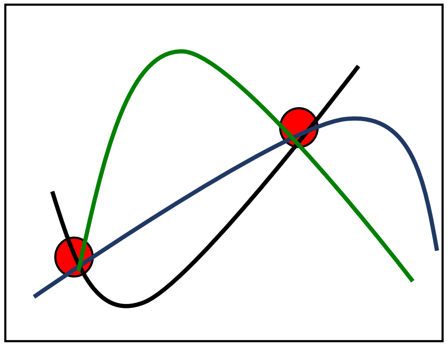


Linear regression

- Linear model, two data



- Quadratic model, two data?
  - Infinitely many settings with zero error
  - How to choose among them?



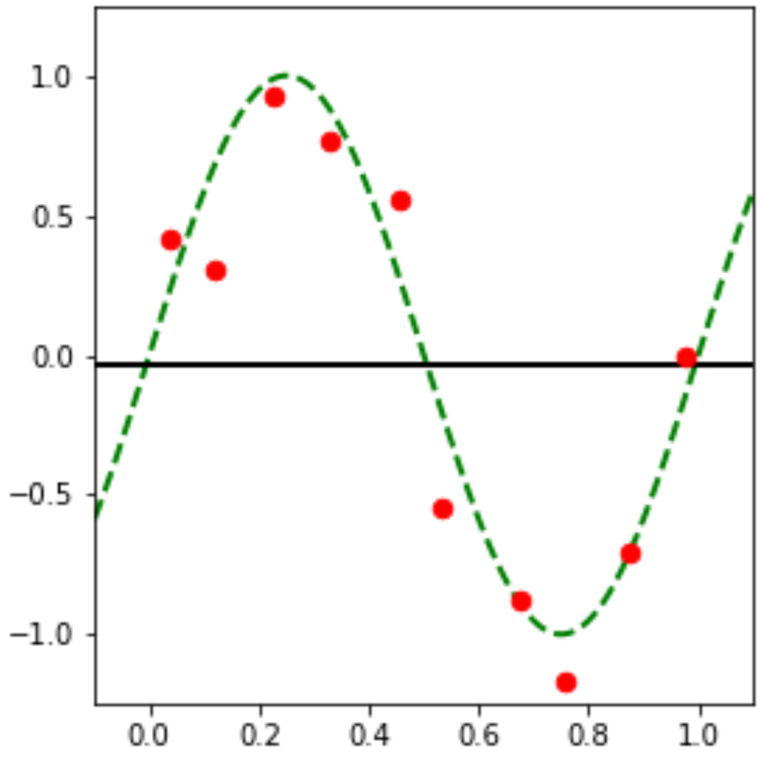
- Higher order coefficents = 0?
  - Uses knowledge of where features came from…
- Could choose e.g. minimum magnitude:

$$\min \underline{\theta}\underline{\theta}^T \quad s.t. \quad J(\underline{\theta}) = 0$$

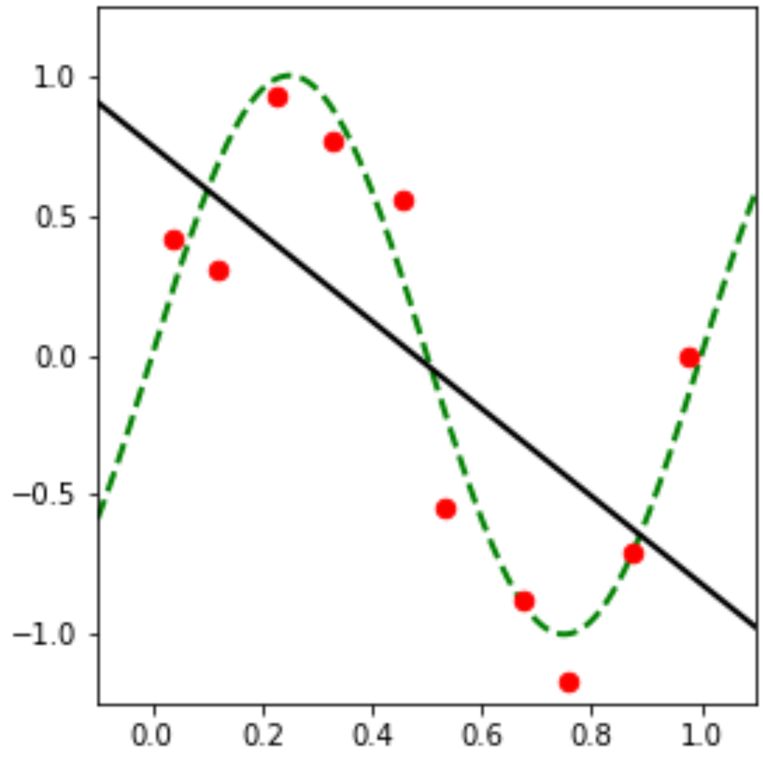
- A type of *bias*: tells us which models to prefer

Sinusoid plus noise

- Fit polynomials:



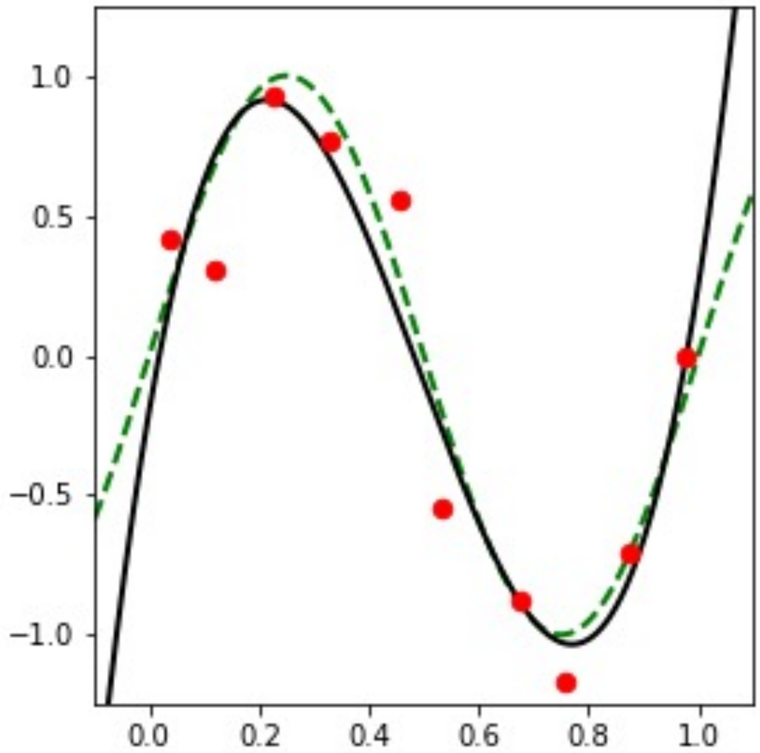
Degree 0
(constant)



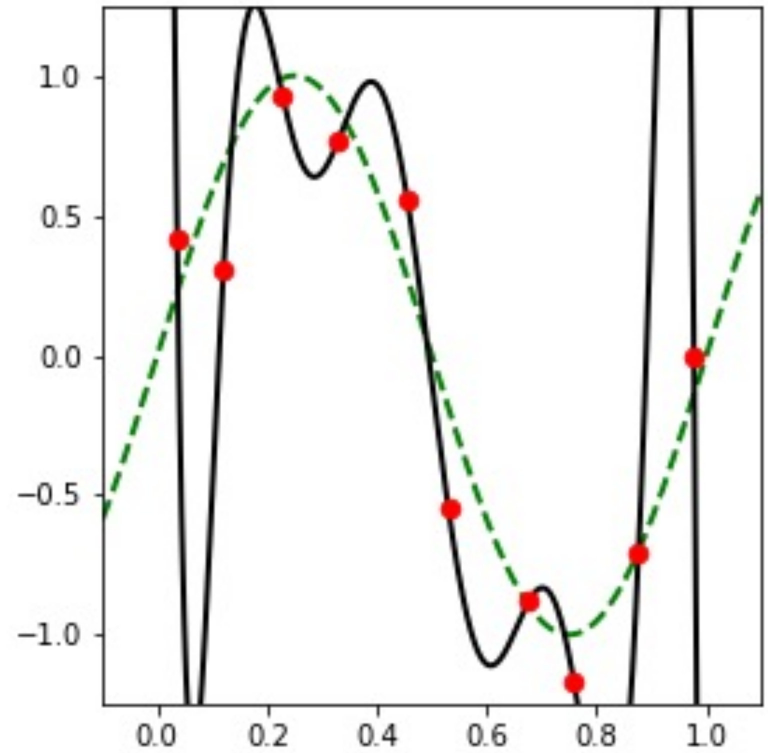
Degree 1
(linear)

Sinusoid plus noise

- Fit polynomials:



Degree 3
(cubic)

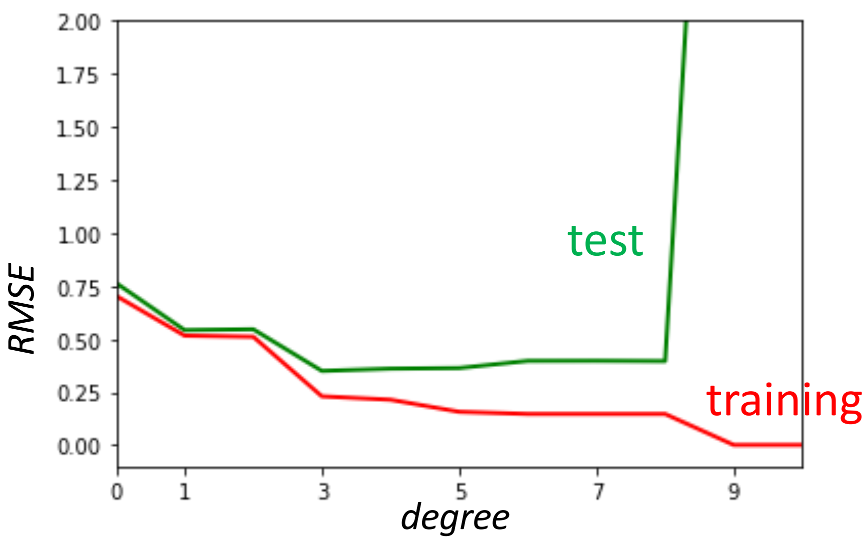


Degree 9

Estimated Polynomial Coefficients

Estimated Regression Coeffients θ

D=0	D=1	D=3	D=9
-0.03	0.75	-0.20	13.63
	-1.72	12.52	-720.38
		-40.89	12766.25
		30.43	-108900.61
			521988.15
			-1502606.92
			2648136.86
			-2794214.56
			1619195.57
			-395947.22



Regularization

- Can modify our cost function J to add "preference" for certain parameter values

$$J(\underline{\theta}) = \frac{1}{2}(\underline{y} - \underline{\theta}\,\underline{X}^T) \cdot (\underline{y} - \underline{\theta}\,\underline{X}^T)^T + \alpha\,\theta\theta^T$$

L_2 penalty:
"Ridge regression"

$$\theta\theta^T = \sum_i \theta_i^2$$

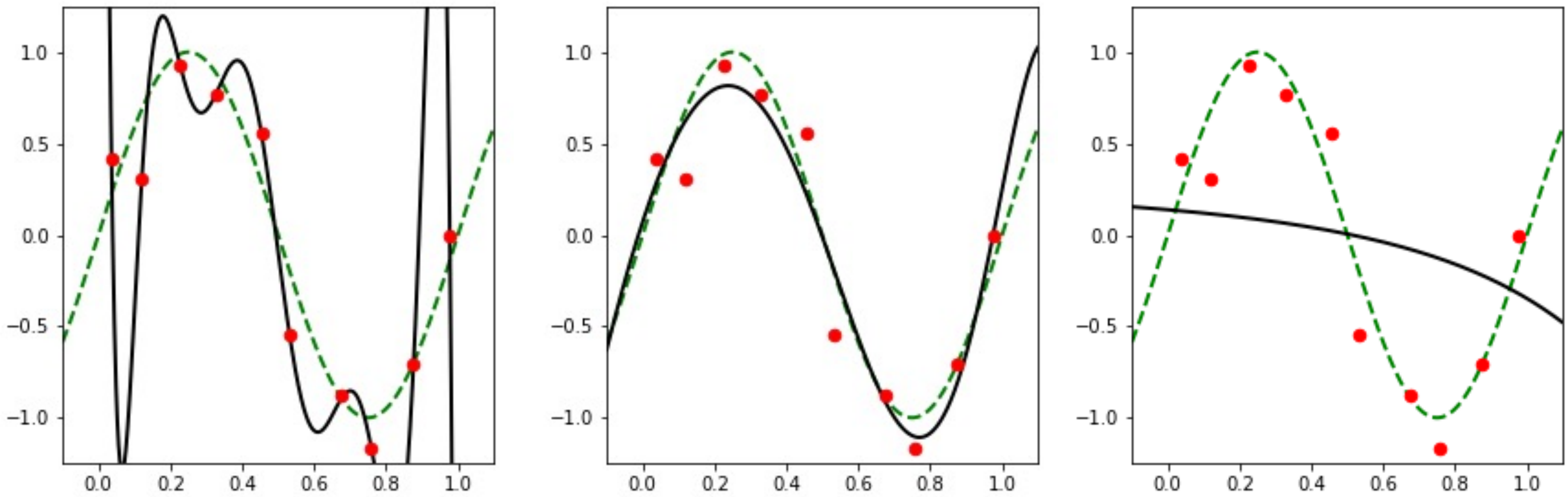
- New solution (derive the same way)

$$\underline{\theta} \quad = \quad \underline{y}\,\underline{X}(\underline{X}^T\,\underline{X} + \alpha I)^{-1}$$

  - Problem is now well-posed for any degree

- Notes:
  - "Shrinks" the parameters toward zero
  - Alpha large: we prefer small theta to small MSE
  - Regularization term is independent of the data: paying more attention reduces our model variance

Regression: Regularization



No regularization

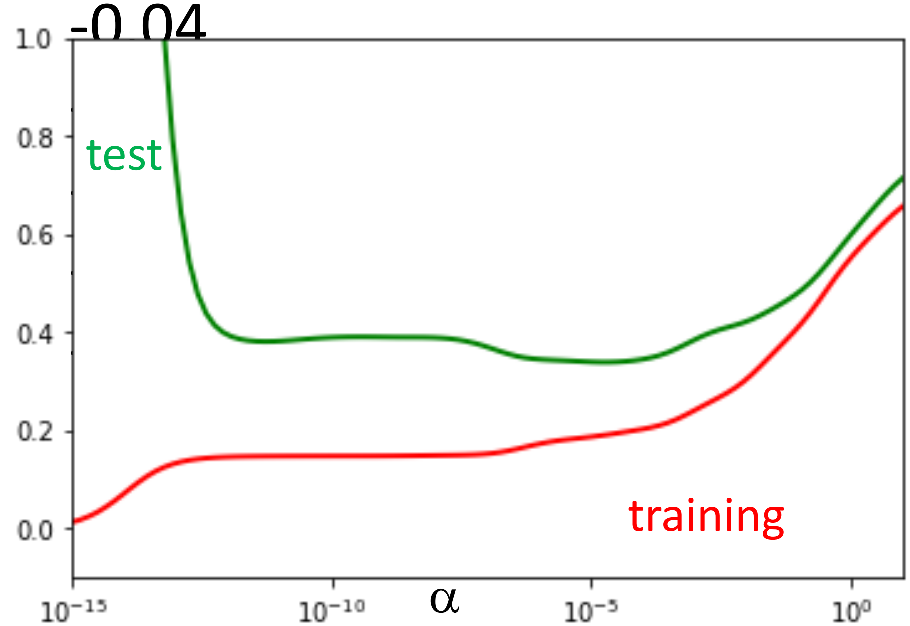
Moderate regularization

Strong regularization

Estimated Polynomial Coefficients

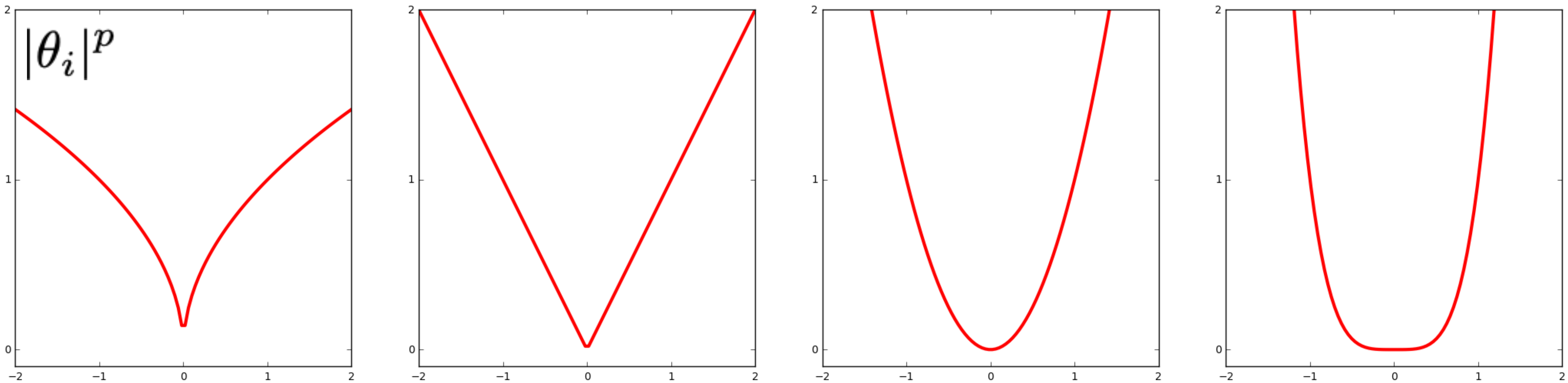
Estimated Regression Coeffients θ

	α zero	α medium	α big
	12.48	0.07	0.14
	-657.38	6.61	-0.19
	11636.48	-12.83	-0.16
	-99177.31	-9.81	-0.11
	475225.43	4.19	-0.07
	-1368006.94	11.12	-0.04
	2411278.80	9.99	
	-2544758.97	4.31	
	1474905.99	-2.87	
	-360722.27	-9.76	



Different regularization functions

- More generally, for the L_p regularizer: $$\Big(\sum_i |\theta_i|^p \Big)^{\frac{1}{p}}$$



p=0.5 p=1 p=2 p=4

Penalize small non-zero parameters more
Prefer some parameters exactly zero;
a few big parameters are OK

Penalize small non-zero parameters less,
but big parameter values a lot
Prefer lots of small weights, no big weights

"Equivalent" formulations

- Many ways to impose the same type of regularization
  - Ex: L2, quadratic penalty on parameter magnitude

$$J(\underline{\theta}) = \frac{1}{2}(\underline{y} - \underline{\theta}\,\underline{X}^T) \cdot (\underline{y} - \underline{\theta}\,\underline{X}^T)^T + \alpha\,\theta\theta^T$$

Scale by data size:

$$\frac{1}{2m}(y - \theta X^T) \cdot (y - \theta X^T)^T + \tilde{\alpha}\theta\theta^T \qquad \tilde{\alpha} = \alpha/m$$

Constraint form:

$$\min_{\theta}(y - \theta X^T) \cdot (y - \theta X^T)^T \quad \text{s.t.} \quad \theta\theta^T \le R$$

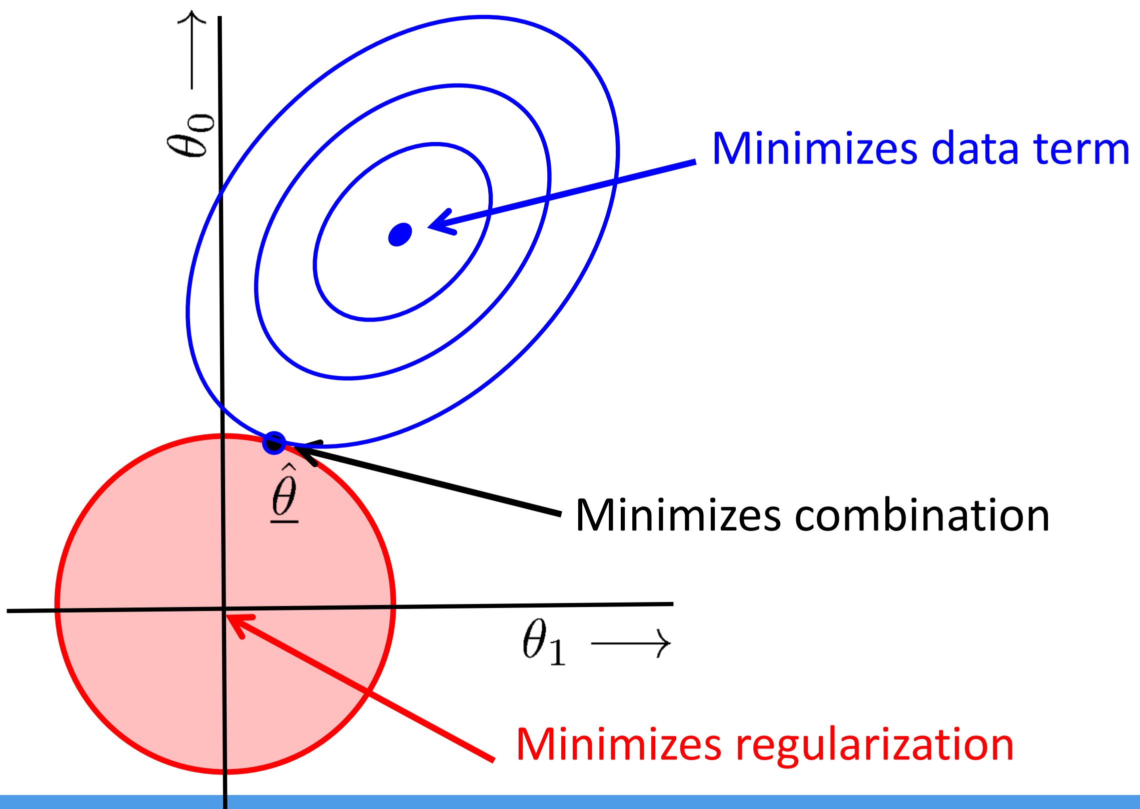
(Lagrangian)

Magnitude vs Squared Magnitude:

$$\frac{1}{2}(y - \theta X^T) \cdot (y - \theta X^T)^T + \gamma(\theta\theta^T)^{\frac{1}{2}} \qquad \gamma = \alpha||\theta^*||$$

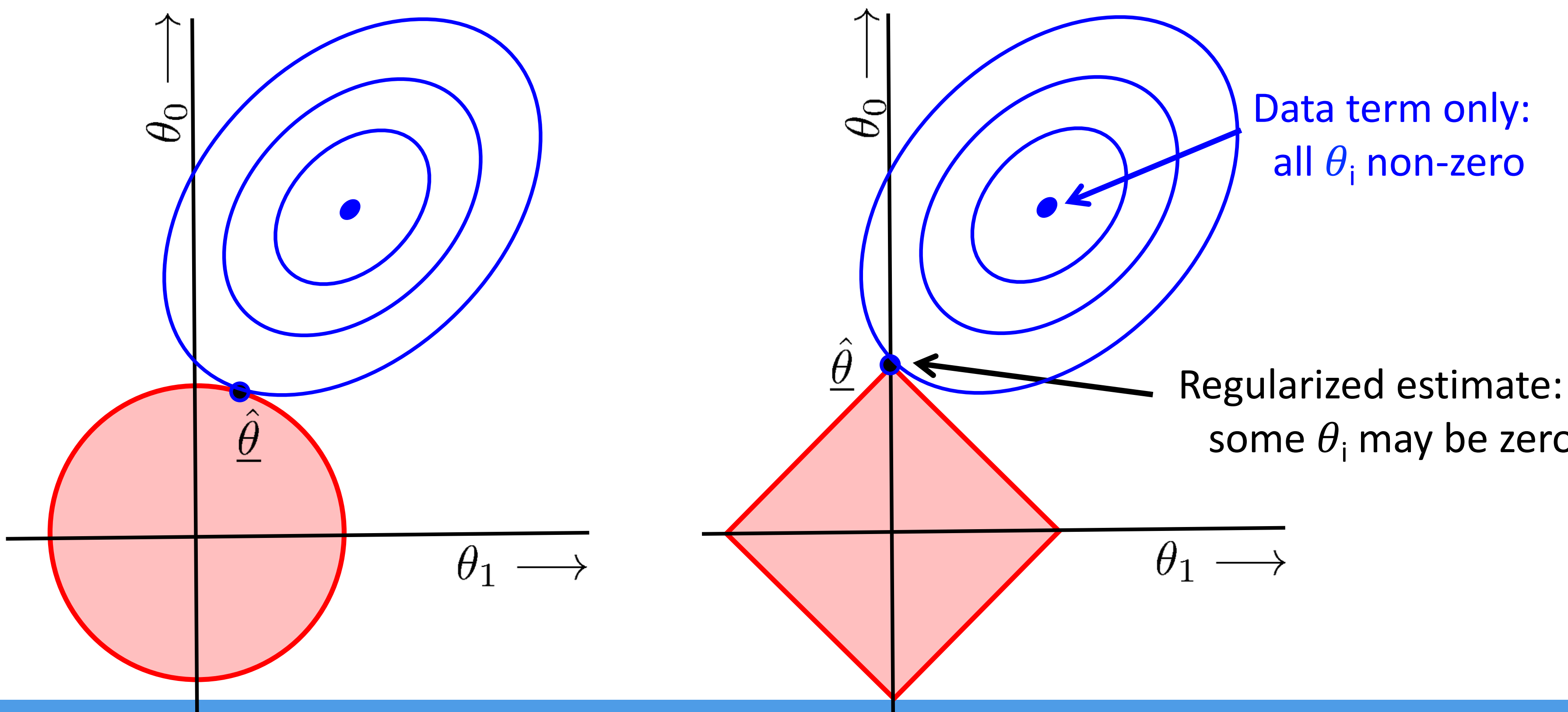
Regularization: L_2 vs L_1

- Estimate balances data term & regularization term



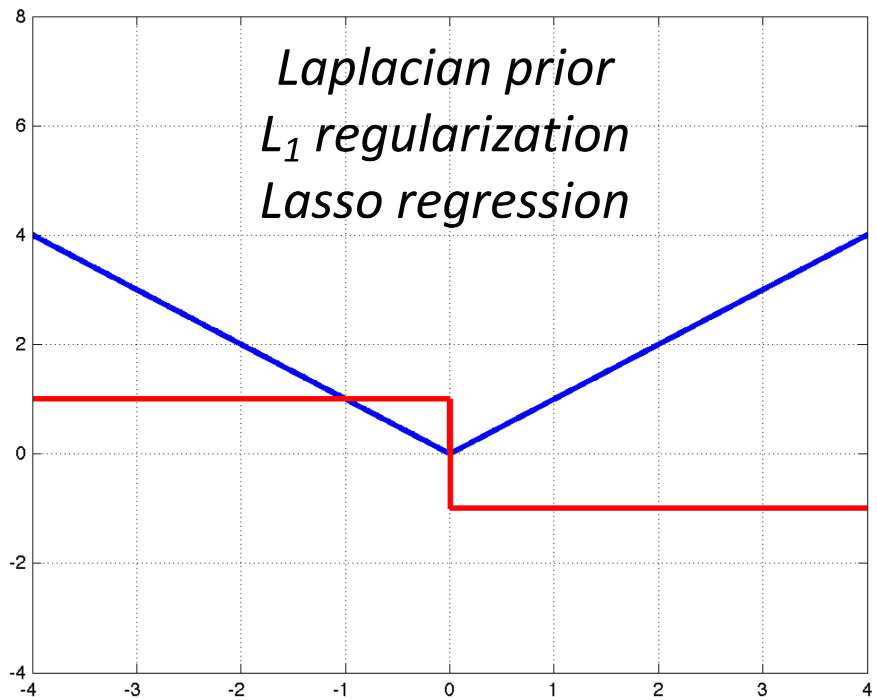
Regularization: L_2 vs L_1

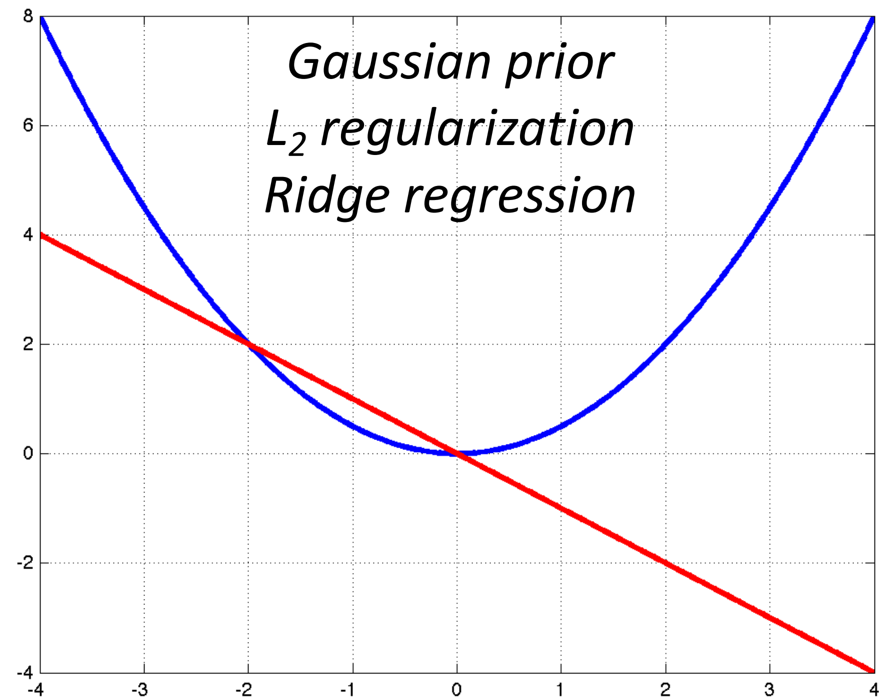
- Estimate balances data term & regularization term
- Lasso tends to generate sparser solutions than a quadratic regularizer.



Gradient-Based Optimization

- L_2 makes (all) coefficients smaller
- L_1 makes (some) coefficients exactly zero: *feature selection*





Objective Function: $f(\theta_i) = |\theta_i|^p$

Negative Gradient: $-f'(\theta_i)$

(Informal intuition: Gradient of L_1 objective not defined at zero)