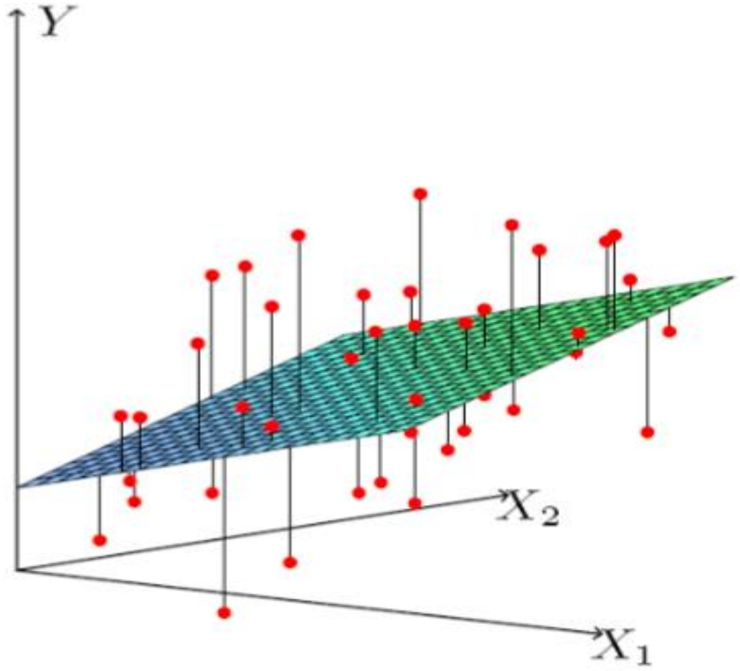
Gradient Descent

Dr. Xiaowei Huang

https://cgi.csc.liv.ac.uk/~xiaowei/

Up to now,

- Three machine learning algorithms:
  - decision tree learning
  - k-nn
  - linear regression
    - linear regression
    - linear classification
    - logistic regression



only optimization objectives are discussed, but how to solve?

Today's Topics

- Derivative
- Gradient
- Directional Derivative
- Method of Gradient Descent
- Example: Gradient Descent on Linear Regression
- Linear Regression: Analytical Solution

Problem Statement: Gradient-Based Optimization

- Most ML algorithms involve optimization
- Minimize/maximize a function *f* (***x***) by altering ***x***
  - Maximization accomplished by minimizing *–f(**x**)*
- *f* (***x***) referred to as objective function or criterion
  - In minimization also referred to as loss function cost, or error
  - Example:
    - linear least squares $f(x) = \frac{1}{2}||Ax - b||^2$
    - Linear regression $\hat{L}(f_w) = \frac{1}{m}\sum_{i=1}^{m}(w^T x^{(i)} - y^{(i)})^2$
- Denote optimum value by ***x****=argmin *f* (***x***)

Derivative

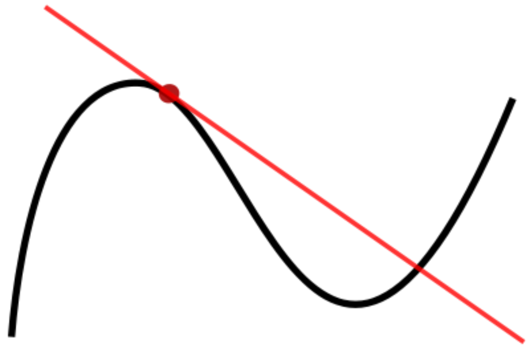
Derivative of a function

- Suppose we have function $y=f(x)$, x, y real numbers
  - Derivative of function denoted: $f'(x)$ or as dy/dx
    - Derivative $f'(x)$ gives the slope of $f(x)$ at point x
    - It specifies how to scale a small change in input to obtain a corresponding change in the output:

$$f(x+\Delta) \approx f(x) + \Delta f'(x)$$

How to design Δ?

    - It tells how you make a small change in input to make a small improvement in y



Recall what's the derivative for the following functions:
$f(x) = x^2$
$f(x) = e^x$
…

Calculus in Optimization

- Suppose we have function $y = f(x)$, where *x, y are* real numbers
- Sign function:

$$sign(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

- We know that

$$f(x - \epsilon sign(f'(x))) < f(x)$$

  for small ε.

This technique is called ***gradient descent*** (Cauchy 1847)

- Therefore, we can reduce $f(x)$ by moving *x* in small steps with opposite sign of derivative

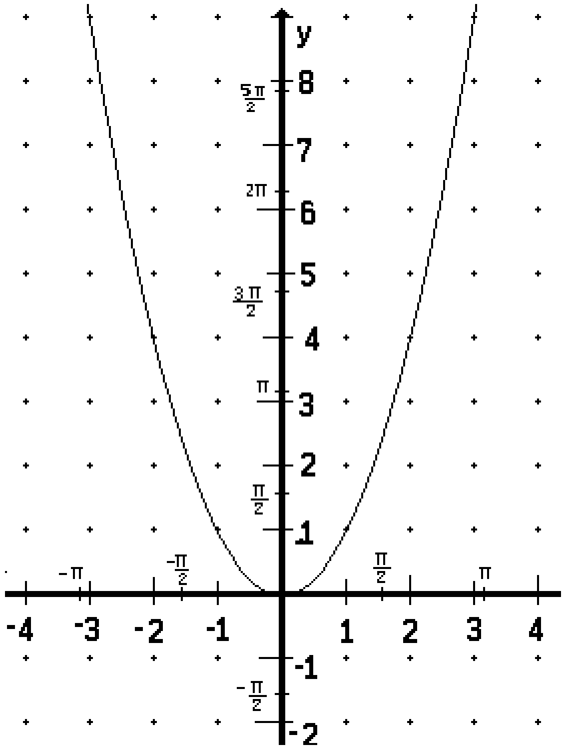
Why opposite?

Example

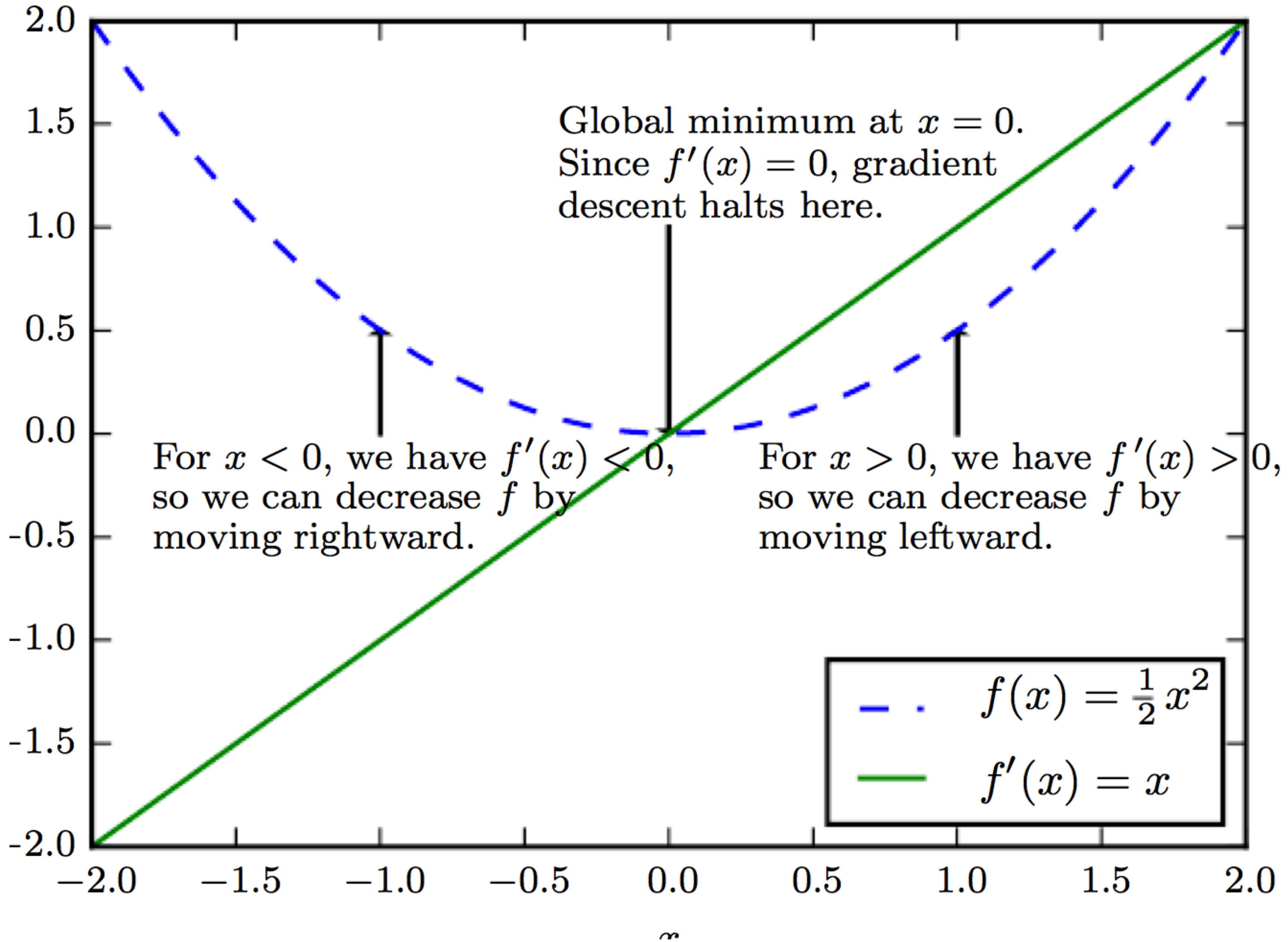
- Function $f(x) = x^2$ $\varepsilon = 0.1$
- $f'(x) = 2x$

- For $x = -2$, $f'(-2) = -4$, $\text{sign}(f'(-2))=-1$
- $f(-2- \varepsilon*(-1)) = f(-1.9) < f(-2)$

- For $x = 2$, $f'(2) = 4$, $\text{sign}(f'(2)) = 1$
- $f(2- \varepsilon*1) = f(1.9) < f(2)$



Gradient Descent Illustrated



For x<0, $f(x)$ decreases with x and *f'(x)<0*

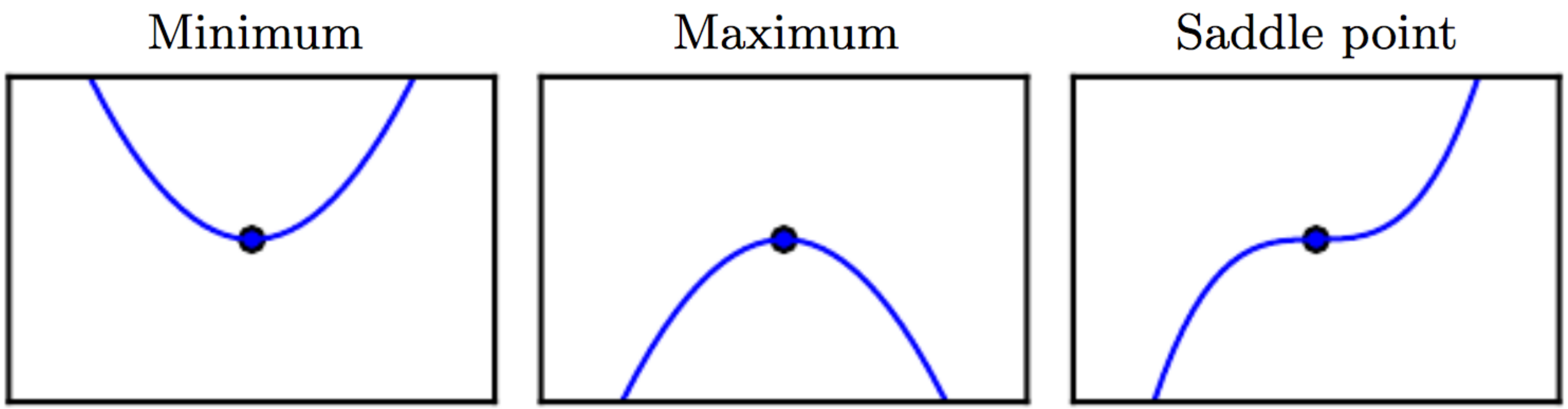
For x>0, *f(x)* increases with x and *f'*(*x*)>*0*

Use *f'(x)* to follow function downhill

Reduce *f(x)* by going in direction opposite sign of derivative *f'*(*x*)

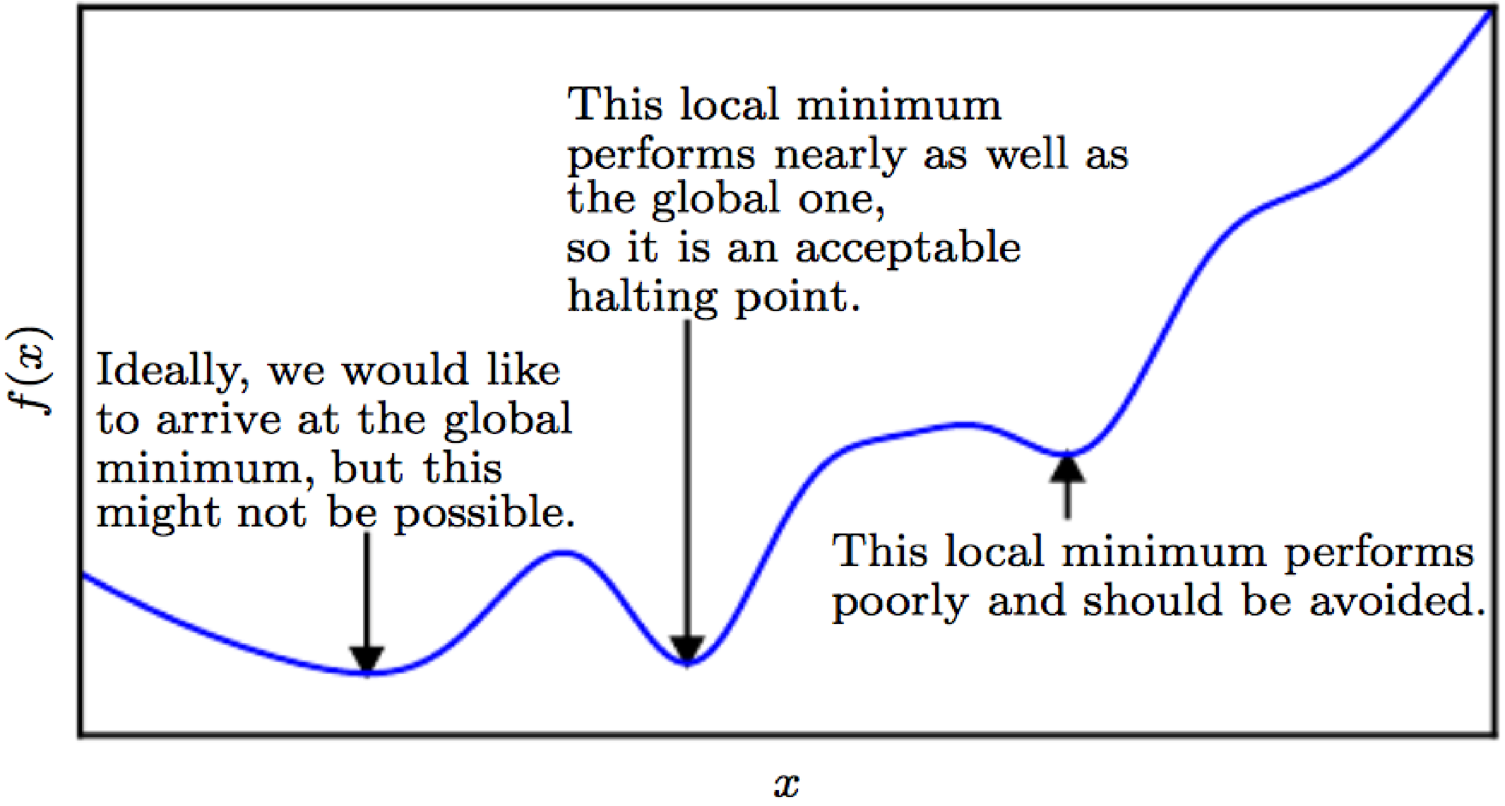
Stationary points, Local Optima

- When $f'(x) = 0$ derivative provides no information about direction of move
- Points where $f'(x) = 0$ are known as *stationary* or critical points
  - Local minimum/maximum: a point where *f(x)* lower/ higher than all its neighbors
  - Saddle Points: neither maxima nor minima



Presence of multiple minima

- Optimization algorithms may fail to find global minimum
- Generally accept such solutions



Gradient

Minimizing with multiple dimensional inputs

- We often minimize functions with multiple-dimensional inputs

$$f : \mathrm{R}^n \rightarrow \mathrm{R}$$

- For minimization to make sense there must still be only one (scalar) output

Functions with multiple inputs

- Partial derivatives

$$\frac{\partial}{\partial x_i} f(x)$$

measures how *f* changes as only variable x_i increases at point ***x***

- Gradient generalizes notion of derivative where derivative is wrt a vector
- Gradient is vector containing all of the partial derivatives denoted

$$\nabla_x f(x) = (\frac{\partial}{\partial x_1} f(x), \ldots, \frac{\partial}{\partial x_n} f(x))$$

Example

- $y = 5x_1^5 + 4x_2 + x_3^2 + 2$
- so what is the exact gradient on instance (1,2,3)?

- the gradient is $(25x_1^4, 4, 2x_3)$
- On the instance (1,2,3), it is (25,4,6)

Functions with multiple inputs

- Gradient is vector containing all of the partial derivatives denoted

$$\nabla_x f(x) = \left(\frac{\partial}{\partial x_1} f(x), \ldots, \frac{\partial}{\partial x_n} f(x)\right)$$

  - Element *i* of the gradient is the partial derivative of *f* wrt x_i
- Critical points are where every element of the gradient is equal to zero

$$\nabla_x f(x) = 0 \equiv \begin{cases} \frac{\partial}{\partial x_1} f(x) = 0 \\ \ldots \\ \frac{\partial}{\partial x_n} f(x) = 0 \end{cases}$$

Example

- $y = 5x_1^5 + 4x_2 + x_3^2 + 2$
- so what are the critical points?

- the gradient is $(25x_1^4, 4, 2x_3)$
- We let $25x_1^4 = 0$ and $2x_3 = 0$, so all instances whose x_1 and x_3 are 0. but 4 /= 0. So there is no critical point.

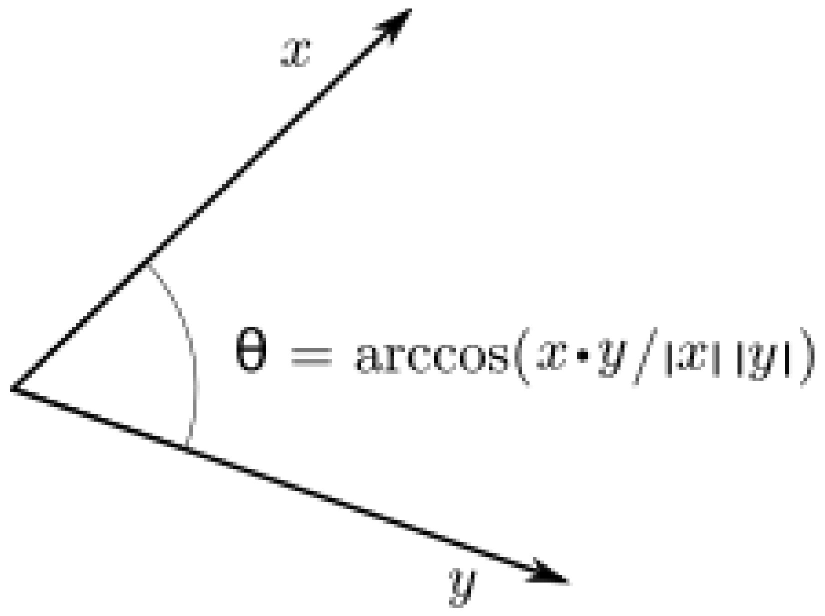
Directional Derivative

Recap: dot product in linear algebra

$$f_w(x) = w^T x$$

$$w = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \qquad x = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

$$w^T x = 2 * 1 + 3 * 4 = 14$$



Geometric meaning: can be used to understand the angle between two vectors

Directional Derivative

- Directional derivative in direction u (a unit vector) is the slope of function f in direction u
- This evaluates to

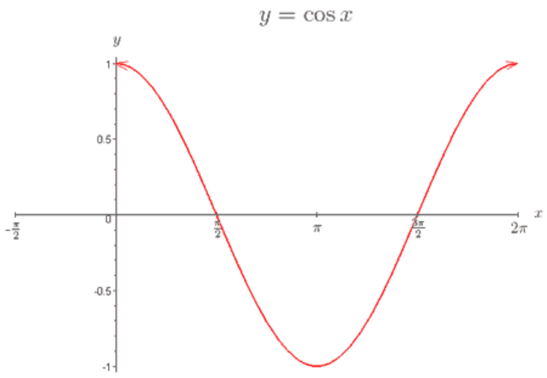
$$u^T \nabla_x f(x)$$

- Example: let $u^T = (u_x, u_y, u_z)$ be a unit vector in Cartesian coordinates, so

$$||u||_2 = \sqrt{u_x^2 + u_y^2 + u_z^2} = 1$$

then

$$u^T \nabla_x f(x) = \frac{\partial f}{\partial x} u_x + \frac{\partial f}{\partial y} u_y + \frac{\partial f}{\partial z} u_z$$



Directional Derivative

- To minimize *f* find direction in which *f* decreases the fastest

$$\min_{u, u^T u = 1} u^T \nabla_x f(x) = \min_{u, u^T u = 1} ||u||_2 \cdot ||\nabla_x f(x)||_2 \cdot \cos \theta$$

  - where θ is angle between u and the gradient
  - Substitute $||u||_2 = 1$ and ignore factors that not depend on u this simplifies to

$$\min_u \cos \theta$$

  - This is minimized when u points in direction opposite to gradient
- In other words, the *gradient points directly uphill, and the negative gradient points directly downhill*

Method of Gradient Descent

Method of Gradient Descent

- The gradient points directly uphill, and the negative gradient points directly downhill
- Thus we can decrease f by moving in the direction of the negative gradient
  - This is known as the method of steepest descent or gradient descent
- Steepest descent proposes a new point

$$x' = x - \epsilon \nabla_x f(x)$$

  - where ϵ is the learning rate, a positive scalar. Set to a small constant.

Choosing ϵ: Line Search

- We can choose ϵ in several different ways
- Popular approach: set ϵ to a small constant
- Another approach is called *line search*:
  - Evaluate

$$f(x - \epsilon \nabla_x f(x))$$

for several values of ϵ and choose the one that results in smallest objective function value

Example: Gradient Descent on Linear Regression

Example: Gradient Descent on Linear Regression

- Linear regression: $\hat{L}(f_w) = \frac{1}{m}\sum_{i=1}^{m}(w^T x^{(i)} - y^{(i)})^2 = \frac{1}{m}||Xw - y||_2^2$

- The gradient is

$$
\begin{aligned}
& \nabla_w \hat{L}(f_w) \\
= \;& \nabla_w \tfrac{1}{m}||Xw - y||_2^2 \\
= \;& \nabla_w [(Xw - y)^T (Xw - y)] \\
= \;& \nabla_w [w^T X^T X w - 2 w^T X^T y + y^T y] \\
= \;& 2X^T X w - 2X^T y
\end{aligned}
$$

Example: Gradient Descent on Linear Regression

- Linear regression: $\hat{L}(f_w) = \frac{1}{m}\sum_{i=1}^{m}(w^T x^{(i)} - y^{(i)})^2 = \frac{1}{m}||Xw - y||_2^2$
- The gradient is $\nabla_w \hat{L}(f_w) = 2X^T Xw - 2X^T y$
- Gradient Descent algorithm is
  - Set step size ϵ, tolerance δ to small, positive numbers.
  - *While* $||X^T Xw - X^T y||_2 > \delta$ do

$$x \longleftarrow x - \epsilon(X^T Xw - X^T y)$$

Linear Regression: Analytical solution

Convergence of Steepest Descent

- Steepest descent converges when every element of the gradient is zero
  - In practice, very close to zero
- We may be able to avoid iterative algorithm and jump to the critical point by solving the following equation for x

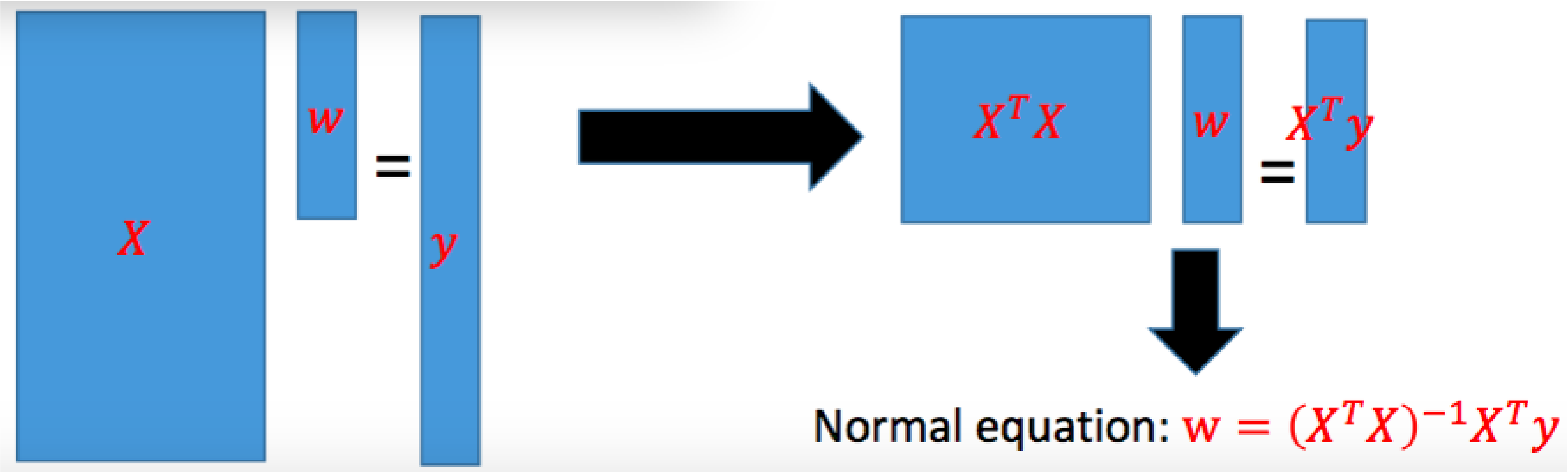
$$\nabla_x f(x) = 0$$

Linear Regression: Analytical solution

- Linear regression: $\hat{L}(f_w) = \frac{1}{m}\sum_{i=1}^{m}(w^T x^{(i)} - y^{(i)})^2 = \frac{1}{m}||Xw - y||_2^2$
- The gradient is $\nabla_w \hat{L}(f_w) = 2X^T X w - 2X^T y$
- Let $\nabla_w \hat{L}(f_w) = 2X^T X w - 2X^T y = 0$
- Then, we have $w = (X^T X)^{-1} X^T y$

Linear Regression: Analytical solution

- Algebraic view of the minimizer
- If X is invertible, just solve $Xw = y$ and get $w = X^{-1}y$
- But typically X is a tall matrix



Generalization to discrete spaces

Generalization to discrete spaces

- Gradient descent is limited to continuous spaces
- Concept of repeatedly making the best small move can be generalized to discrete spaces
- Ascending an objective function of discrete parameters is called *hill climbing*

Exercises

- Given a function $f(x)= e^x/(1+e^x)$, how many critical points?
- Given a function $f(x_1,x_2)= 9x_1^2+3x_2+4$, how many critical points?
- Please write a program to do the following: given any differentiable function (such as the above two), an ε, and a starting x and a target x', determine whether it is possible to reach x' from x. If possible, how many steps? You can adjust ε to see the change of the answer.