CPSC 340:
Machine Learning and Data Mining

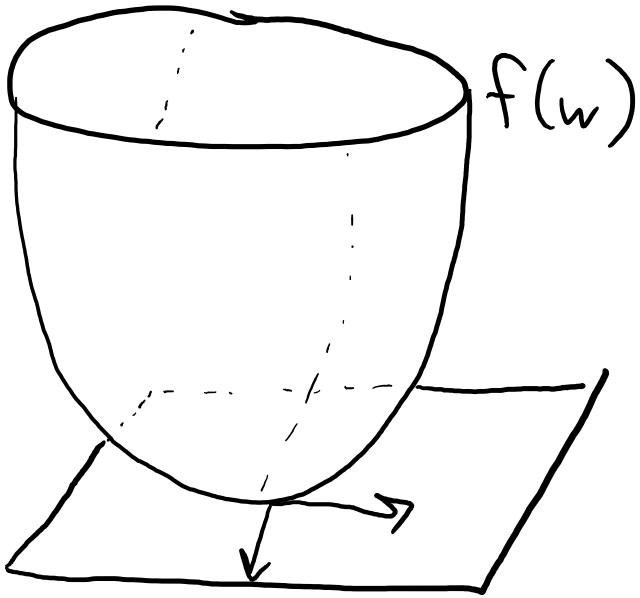
The Normal Equations

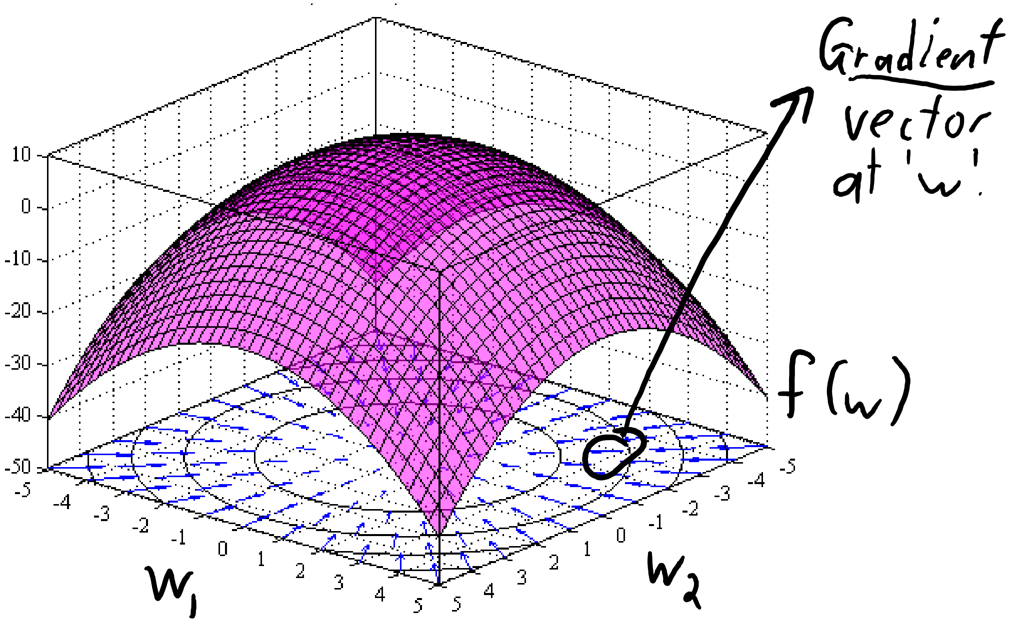
Fall 2017

Gradient and Critical Points in d-Dimensions

- Generalizing “set the derivative to 0 and solve” in d-dimensions:
  - Find ‘w’ where the gradient vector equals the zero vector.
- Gradient is vector with partial derivative ‘j’ in position ‘j’:

$$\nabla f(w) = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \vdots \\ \frac{\partial f}{\partial w_d} \end{bmatrix}$$





Tangent slope is 0 in every direction at minimizer.

ac.redwoods.edu/~darnold/math50c/matlab/pderiv/index.xhtml

Least Squares Partial Derivatives

- The linear least squares model in d-dimensions minimizes:

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(w^Tx_i - y_i)^2$$

$w^Tx_i = w_1x_{i1} + w_2x_{i2} + \cdots + w_dx_{id}$

$\frac{d}{dw_1}[w^Tx_i] = x_{i1} + 0 + \cdots + 0$

$= x_{i1}$

- Computing the partial derivative:

$$\frac{\partial}{\partial w_1}\left[\frac{1}{2}\sum_{i=1}^{n}(w^Tx_i - y_i)^2\right] = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial w_1}\left[(w^Tx_i - y_i)^2\right]$$

$$= \frac{1}{2}\sum_{i=1}^{n} 2(w^Tx_i - y_i)\underbrace{\frac{\partial}{\partial w_1}[w^Tx_i]}_{\text{What is the derivative of } w^Tx_i \text{ with respect to } w_1?}$$

$$= \sum_{i=1}^{n}(w^Tx_i - y_i)x_{i1}$$

Problem: I can't just set to 0 and solve because it depends on $w_2, w_3, \ldots, w_d$

Gradient and Critical Points in d-Dimensions

- Generalizing "set the derivative to 0 and solve" in d-dimensions:
  - Find 'w' where the gradient vector equals the zero vector.
- Gradient is vector with partial derivative 'j' in position 'j':

$$\nabla f(w) = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \vdots \\ \frac{\partial f}{\partial w_d} \end{bmatrix}$$

For linear least squares:

$$\nabla f(w) = \begin{bmatrix} \sum_{i=1}^{n} (w^T x_i - y_i) x_{i1} \\ \sum_{i=1}^{n} (w^T x_i - y_i) x_{i2} \\ \vdots \\ \sum_{i=1}^{n} (w^T x_i - y_i) x_{id} \end{bmatrix}$$

Claims for linear least square:

1. Finding a 'w' where $\nabla f(w) = 0$ can be done by solving a <u>system of linear equations</u>.
2. <u>All</u> 'w' where $\nabla f(w) = 0$ are <u>minimizers</u>.

ac.redwoods.edu/~darnold/math50c/matlab/pderiv/index.xhtml

Matrix/Norm Notation (MEMORIZE/STUDY THIS)

- To solve the d-dimensional least squares, we use matrix notation:
  - We use ‘y’ as an “n times 1” vector containing target ‘y_i’ in position ‘i’.
  - We use ‘x_i’ as a “d times 1” vector containing features ‘j’ of example ‘i’.
    - We’re now going to be careful to make sure these are column vectors.
  - So ‘X’ is a matrix with the x_i^T in row ‘i’.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} = \begin{bmatrix} — & x_1^T & — \\ — & x_2^T & — \\ & \vdots & \\ — & x_n^T & — \end{bmatrix}$$

Matrix/Norm Notation (MEMORIZE/STUDY THIS)

- To solve the d-dimensional least squares, we use matrix notation:
  - Our prediction for example 'i' is given by scalar w^Tx_i.
  - The matrix-vector product Xw gives predictions for all 'i' (n times 1 vector).

$$w^T x_i = \sum_{j=1}^{d} w_j x_{ij}$$

$$= w_1 x_{i1} + w_2 x_{i2} + \cdots + w_d x_{id}$$

Also, because $w^T x_i$ is a scalar, we have $w^T x_i = x_i^T w$

(e.g., $[5]^T = [5]$)

$$Xw = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} x_{11}w_1 + x_{12}w_2 + \cdots + x_{1d}w_d \\ x_{21}w_1 + x_{22}w_2 + \cdots + x_{2d}w_d \\ \vdots \\ x_{n1}w_1 + x_{22}w_2 + \cdots + x_{nd}w_d \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{d} x_{1j}w_j \\ \sum_{j=1}^{d} x_{2j}w_j \\ \vdots \\ \sum_{j=1}^{d} x_{nj}w_j \end{bmatrix} = \begin{bmatrix} x_1^T w \\ x_2^T w \\ \vdots \\ x_n^T w \end{bmatrix} = \begin{bmatrix} w^T x_1 \\ w^T x_2 \\ \vdots \\ w^T x_d \end{bmatrix}$$

Prediction for example 'i' is in row 'i'

Matrix/Norm Notation (MEMORIZE/STUDY THIS)

- To solve the d-dimensional least squares, we use matrix notation:
  - Our prediction for example ‘i’ is given by scalar w^Tx_i.
  - The matrix-vector product Xw gives predictions for all ‘i’ (n times 1 vector).
  - The residual vector r gives w^Tx_i minus y_i for all ‘i’ (n times 1 vector).
  - Least squares can be written as the squared L2-norm of the residual.

$$r = \begin{bmatrix} w^Tx_1 - y_1 \\ w^Tx_2 - y_2 \\ \vdots \\ w^Tx_n - y_n \end{bmatrix} = \underbrace{\begin{bmatrix} w^Tx_1 \\ w^Tx_2 \\ \vdots \\ w^Tx_n \end{bmatrix}}_{Xw} - \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{y} = Xw - y$$

$$\sum_{i=1}^n (w^Tx_i - y_i)^2 = \sum_{i=1}^n (r_i)^2 = \sum_{i=1}^n r_i r_i = r^Tr = \|r\|^2 = \|Xw - y\|^2$$

Matrix Algebra Review (MEMORIZE/STUDY THIS)

- Review of linear algebra operations we'll use:
  - If 'a' and 'b' be vectors, and 'A' and 'B' be matrices then:

$$a^T b = b^T a$$

$$\|a\|^2 = a^T a$$

$$(A+B)^T = A^T + B^T$$

$$(AB)^T = B^T A^T$$

$$(A+B)(A+B) = AA + BA + AB + BB$$

$$a^T \underbrace{Ab}_{\text{vector}} = \underbrace{b^T A^T}_{\text{vector}} a$$

Sanity check:

ALWAYS CHECK THAT DIMENSIONS MATCH

(if not, you did something wrong)

Linear Least Squares

Want 'w' that minimizes

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(w^Tx_i - y_i)^2 = \underbrace{\frac{1}{2}\|Xw - y\|_2^2}_{\text{Let's expand then compute gradient.}} = \frac{1}{2}(Xw-y)^T(Xw-y)$$

$$= \frac{1}{2}((Xw)^T - y^T)(Xw-y)$$

$$= \frac{1}{2}(w^TX^T - y^T)(Xw-y)$$

$$= \frac{1}{2}(w^TX^T(Xw-y) - y^T(Xw-y))$$

$$= \frac{1}{2}(w^TX^TXw - w^TX^Ty - y^TXw + y^Ty)$$

$$= \underbrace{\frac{1}{2}w^TX^TXw - w^TX^Ty + \frac{1}{2}y^Ty}_{\text{Sanity check: all of these are scalars.}}$$

Linear and Quadratic Gradients

- We've written as a d-dimensional quadratic:

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(w^Tx_i - y_i)^2 = \frac{1}{2}\|Xw - y\|^2 = \frac{1}{2}w^T\underbrace{X^TX}_{\text{matrix 'A'}}w - w^T\underbrace{X^Ty}_{\text{vector 'b'}} + \underbrace{\frac{1}{2}y^Ty}_{\text{scalar 'c'}}$$

$$= \frac{1}{2}w^TAw + w^Tb + c$$

- How do we compute gradient?

Let's first do it with $d=1$:

$$f(w) = \frac{1}{2}waw + wb + c$$

$$= \frac{1}{2}aw^2 + wb + c$$

$$f'(w) = aw + b + 0$$

Here are the generalizations to 'd' dimensions:

$$\nabla[c] = 0 \quad \text{(zero vector)}$$

$$\nabla[w^Tb] = b$$

$$\nabla[\tfrac{1}{2}w^TAw] = Aw \quad \text{(if } A \text{ is } \underline{\text{symmetric}}\text{)}$$

Full derivations are on webpage in notes on linear and quadratic gradients.

Linear and Quadratic Gradients

- We've written as a d-dimensional quadratic:

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(w^Tx_i - y_i)^2 = \frac{1}{2}\|Xw - y\|^2 = \frac{1}{2}w^T\underbrace{X^TX}_{\text{matrix 'A'}}w - w^T\underbrace{X^Ty}_{\text{vector 'b'}} + \underbrace{\frac{1}{2}y^Ty}_{\text{scalar 'c'}}$$

$$= \frac{1}{2}w^TAw + w^Tb + c$$

- Gradient is given by: $\nabla f(w) = Aw + b + 0$

- Using definitions of 'A' and 'b': $= \underbrace{X^TXw - X^Ty}_{\text{Sanity check: these are both } d\times 1 \text{ vectors.}}$

Normal Equations

- Set gradient equal to zero to find the "critical" points:

$$X^T X w - X^T y = 0$$

- We now move terms not involving 'w' to the other side:

$$X^T X w = X^T y$$

- This is a set of 'd' linear equations called the normal equations.
  - This a linear system like "Ax = b" from Math 152.
    - You can use Gaussian elimination to solve for 'w'.
  - In Julia, the "\" command can be used to solve linear systems:

Train: w=(X'X)\(X'y)   Predict: yhat = X_{test}*w

Incorrect Solutions to Least Squares Problem

The least squares objective is $f(w) = \frac{1}{2}\|Xw - y\|^2$

The minimizers of this objective are solutions to the linear system:

$$X^T X w = X^T y$$

The following are not the solutions to the least squares problem:

$w = (X^T X)^{-1}(X^T y)$ (only true if $X^T X$ is invertible)

$w X^T X = X^T y$ (matrix multiplication is not commutative, dimensions don't even match)

$w = \frac{X^T y}{X^T X}$ (you cannot divide by a matrix)

Least Squares Issues

- Issues with least squares model:
  - Solution might not be unique.
  - It is sensitive to outliers.
  - It always uses all features.
  - Data can might so big we can't store X^TX.
  - It might predict outside range of y_i values.
  - It assumes a linear relationship between x_i and y_i.

X is $n \times d$

so X^T is $d \times n$

and X^TX is $d \times d$.

Costs $O(nd^2)$ to calculate:

- Each of the $O(d^2)$ elements is an inner product between length 'n' vectors.

Non-Uniqueness of Least Squares Solution

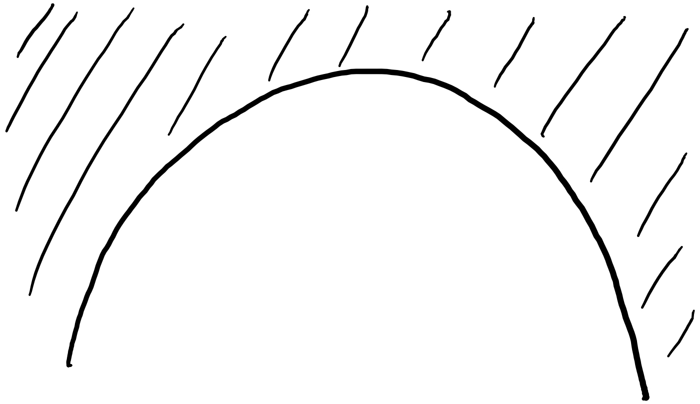
- Why isn't solution unique?
  - Imagine have two features that are identical for all examples.
  - This is special case of features being "collinear"
    - One feature is a linear function of another.
  - I can increase weight on one feature, and decrease it on the other, without changing predictions.

$$\hat{y}_i = w_1 x_{i1} + w_2 \underbrace{x_{i1}}_{\text{copy}} = (w_1 + w_2) x_{i1} + 0 x_{i1}$$

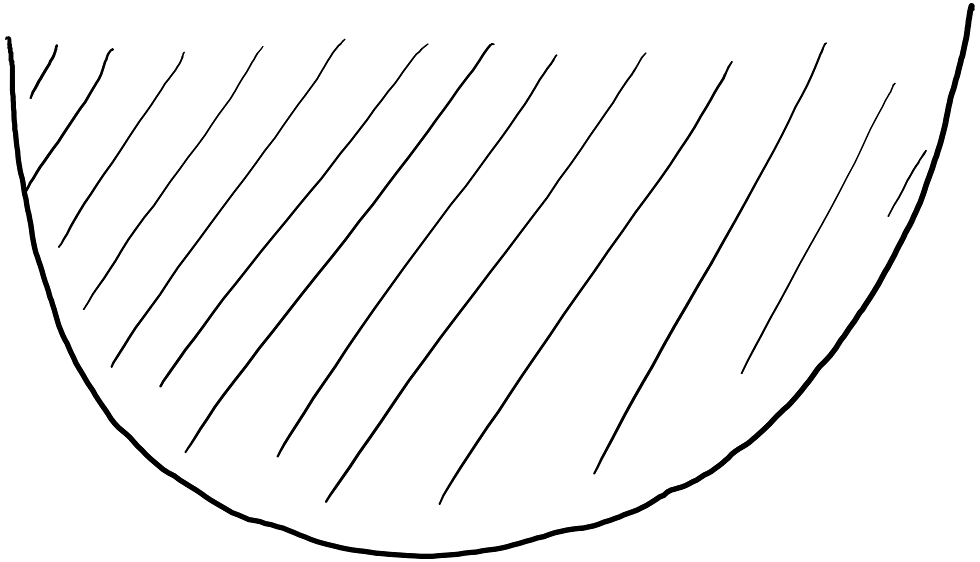
  - Thus, if (w_1,w_2) is a solution then $(w_1+w_2, 0)$ is a solution.

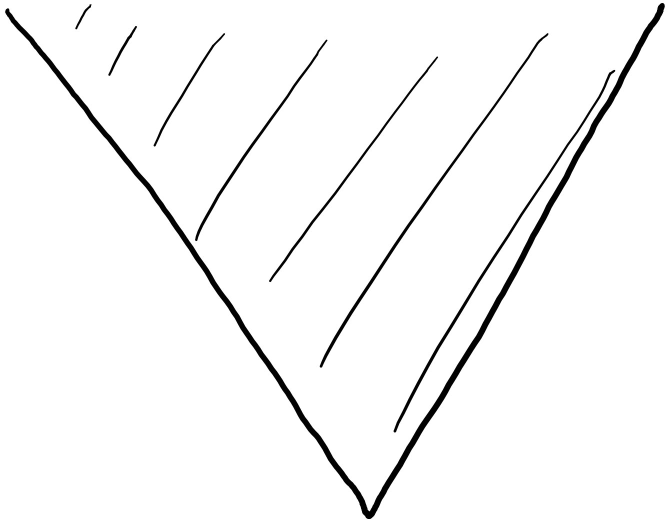
- But, any 'w' where $\nabla f(w) = 0$ is a global optimum, due to convexity.

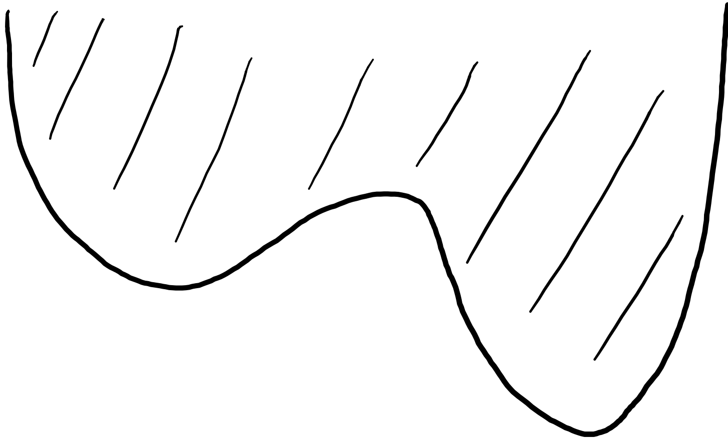
Convex Functions



- Is finding a 'w' with ∇f(w) = 0 good enough?
  - Yes, for convex functions.







- A function is convex if the area above the function is a convex set.
  - All values between any two points above function stay above function.

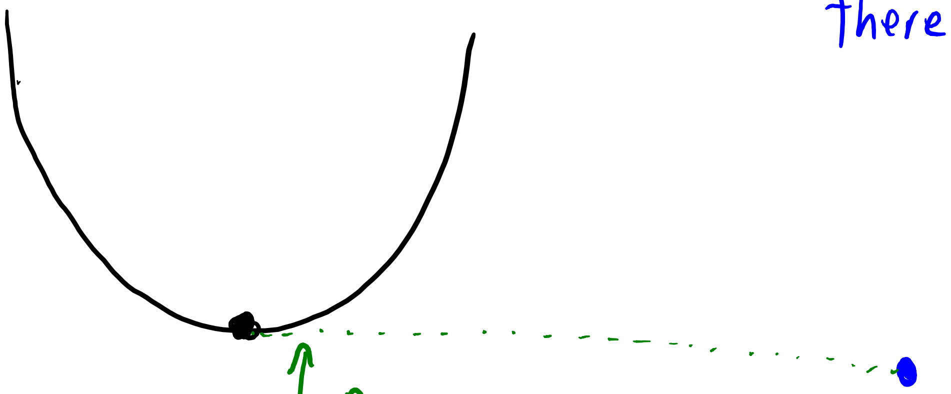
Convex Functions

- All ‘w’ with ∇ f(w) = 0 for convex functions are global minima.

Proof by contradiction:

Consider a local minimum

If this is not global minimum, there must a smaller value.



By convexity we can move along line to global minimum and decrease objective.

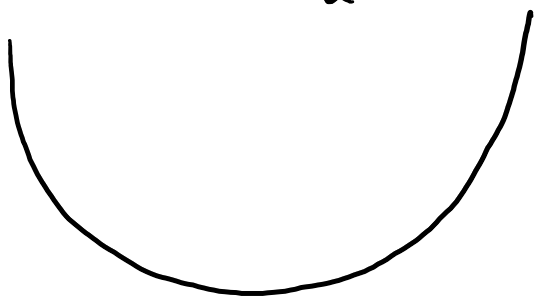
But this contradicts that we are at a local minimum.

  - Normal equations finds a global minimum because of convexity.

How do we know if a function is convex?

- Some useful tricks for showing a function is convex:
  - 1-variable, twice-differentiable function is convex iff $f''(w) \geq 0$ for all 'w'.

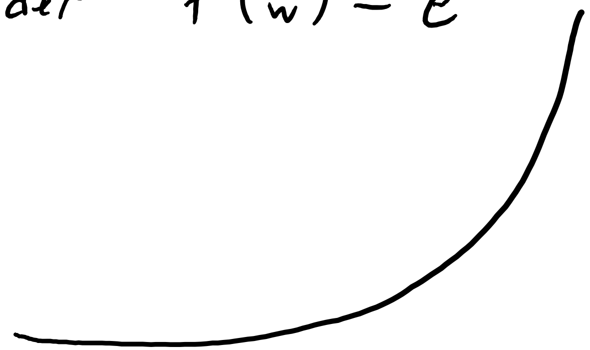
Consider $f(w) = \frac{1}{2}aw^2$ for $a > 0$.



We have $f'(w) = aw$
and $f''(w) = a > 0$

By assumption

Consider $f(w) = e^w$



We have $f'(w) = e^w$
and $f''(w) = e^w > 0$

By definition of exponential function.

How do we know if a function is convex?

- Some useful tricks for showing a function is convex:
  - 1-variable, twice-differentiable function is convex iff $f''(w) \geq 0$ for all ‘w’.
  - A convex function multiplied by non-negative constant is convex.

We showed that $f(w) = e^w$ is convex, so $f(w) = 10e^w$ is convex.

How do we know if a function is convex?

- Some useful tricks for showing a function is convex:
  - 1-variable, twice-differentiable function is convex iff $f''(w) \geq 0$ for all 'w'.
  - A convex function multiplied by non-negative constant is convex.
  - Norms and squared norms are convex.

$\|w\|$, $\|w\|^2$, $\|w\|_1$, $\|w\|_\infty$, $\|w\|_1^2$, and so on are all convex.

How do we know if a function is convex?

- Some useful tricks for showing a function is convex:
  - 1-variable, twice-differentiable function is convex iff $f''(w) \geq 0$ for all ‘w’.
  - A convex function multiplied by non-negative constant is convex.
  - Norms and squared norms are convex.
  - The sum of convex functions is a convex function.

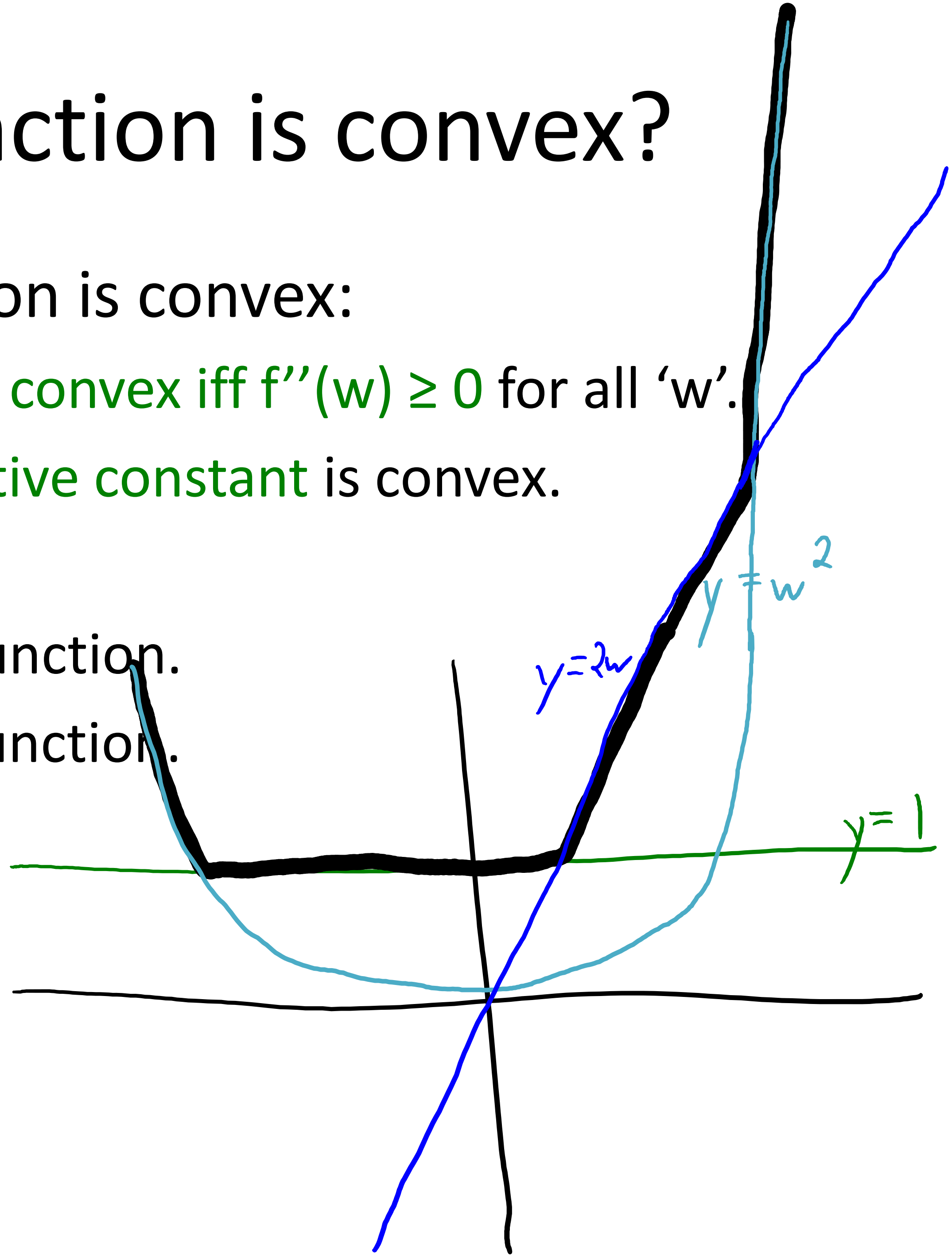
$$f(x) = \underbrace{10e^{w}}_{\text{From earlier}} + \underbrace{\tfrac{1}{2}}_{\text{constant}}\underbrace{\|w\|^2}_{\text{norm squared}} \quad \text{is convex}$$

How do we know if a function is convex?

- Some useful tricks for showing a function is convex:
  - 1-variable, twice-differentiable function is convex iff $f''(w) \geq 0$ for all 'w'.
  - A convex function multiplied by non-negative constant is convex.
  - Norms and squared norms are convex.
  - The sum of convex functions is a convex function.
  - The max of convex functions is a convex function.

$f(w) = \max\{1, 2w, w^2\}$ is convex.

convex



How do we know if a function is convex?

- Some useful tricks for showing a function is convex:
  - 1-variable, twice-differentiable function is convex iff $f''(w) \geq 0$ for all 'w'.
  - A convex function multiplied by non-negative constant is convex.
  - Norms and squared norms are convex.
  - The sum of convex functions is a convex function.
  - The max of convex functions is a convex function.
  - Composition of a convex function and a linear function is convex.

If 'f' is convex the $f(\underbrace{Xw - y}_{\text{linear function}})$ is convex.

How do we know if a function is convex?

- Some useful tricks for showing a function is convex:
  - 1-variable, twice-differentiable function is convex iff $f''(w) \geq 0$ for all 'w'.
  - A convex function multiplied by non-negative constant is convex.
  - Norms and squared norms are convex.
  - The sum of convex functions is a convex function.
  - The max of convex functions is a convex function.
  - Composition of a convex function and a linear function is convex.
- But: not true that composition of convex with convex is convex:

Even if 'f' is convex and 'g' is convex, $f(g(w))$ might <u>not</u> be convex.

E.g. x^2 is convex and $-\log(x)$ is convex but $-\log(x^2)$ is <u>not</u> convex.

Example: Convexity of Linear Regression

- Consider linear regression objective with squared error:

$$f(w) = \|Xw - y\|^2$$

- We can use that this is a convex function composed with linear:

Let $g(r) = \|r\|^2$, which is convex because it's a squared norm.

Then $f(w) = g(Xw - y)$, which is convex because it's a convex function composed with the linear function $h(w) = Xw - y$.

Summary

- Normal equations: solution of least squares as a linear system.
  - Solve $(X^TX)w = (X^Ty)$.
- Solution might not be unique because of collinearity.
- But any solution is optimal because of convexity.
- Convex functions:
  - Set of functions with property that $\nabla f(w) = 0$ implies 'w' is a global min.
  - Can (usually) be identified using a few simple rules.

- Next time: overview of numerical optimization concepts.

Convexity, min, and argmin

- If a function is convex, then all stationary points are global optima.

- However, convex functions don't necessarily have stationary points:
  - For example, f(x) = a*x, f(x) = exp(x), etc.

- Also, more than one 'x' can achieve the global optimum:
  - For example, f(x) = c is minimized by any 'x'.

Bonus Slide: Householder(-ish) Notation

- Househoulder notation: set of (fairly-logical) conventions for math.

Use greek letters for scalars: $\alpha = 1$, $\beta = 3.5$, $\gamma = \pi$

Use first/last lowercase letters for vectors: $w = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}$, $x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $y = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$, $a = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$, $b = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$

→ Assumed to be column-vectors.

Use first/last uppercase letters for matrices: X, Y, W, A, B

Indices use i, j, k.

Sizes use $m, n, d, p,$ and k

hopefully meaning of 'k' is obvious from context

Sets use S, T, U, V

Functions use $f, g,$ and h.

When I write x_i I mean "grab row 'i' of X and make a column-vector with its values."

Bonus Slide: Householder(-ish) Notation

- Househoulder notation: set of (fairly-logical) conventions for math:

Our ultimate least squares notation:

$$f(w) = \frac{1}{2}\|Xw - y\|^2$$

But if we agree on notation we can quickly understand:

$$g(x) = \frac{1}{2}\|Ax - b\|^2$$

If we use random notation we get things like:

$$H(\beta) = \frac{1}{2}\|R\beta - P_n\|^2$$

Is this the same model?

When does least squares have a unique solution?

- We said that least squares solution is not unique if we have repeated columns.
- But there are other ways it could be non-unique:
  - One column is a scaled version of another column.
  - One column could be the sum of 2 other columns.
  - One column could be three times one column minus four times another.

- Least squares solution is unique if and only if all columns of X are "linearly independent".
  - No column can be written as a "linear combination" of the others.
  - Many equivalent conditions (see Strang's linear algebra book):
    - X has "full column rank", X^TX is invertible, X^TX has non-zero eigenvalues, $\det(X^TX) > 0$.
  - Note that we cannot have independent columns if $d > n$.