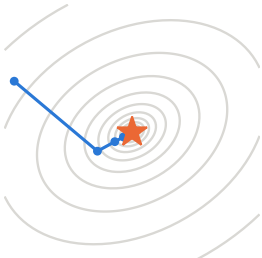
# Newton’s method

**208424 Lecture 5**

# Lecture 4: ill-conditioning is slow

$\kappa = 2$: 20 steps



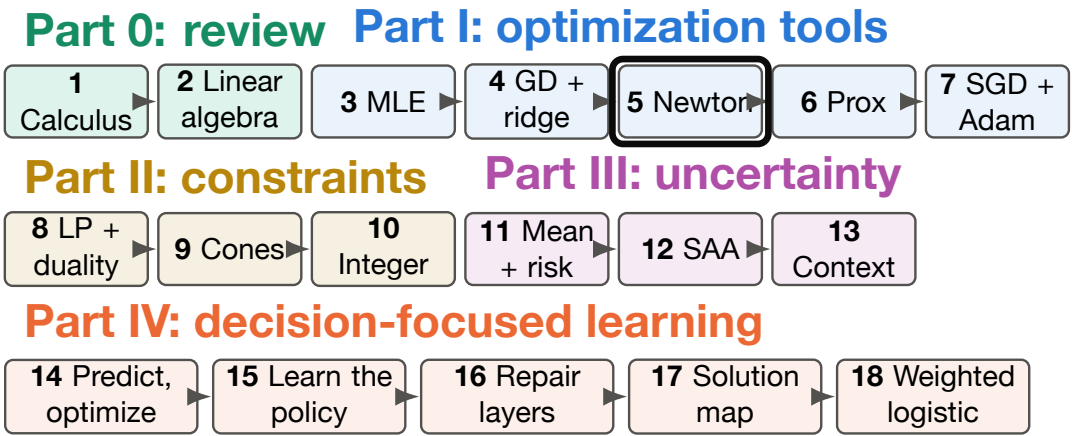
$\kappa = 25$: 327 steps



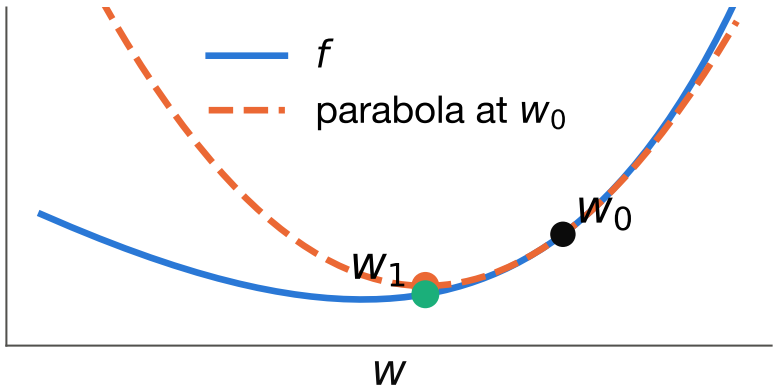
Lecture 4: gradient descent uses only the slope.
20 steps when well-conditioned, 327 when ill-conditioned.

today: use the curvature too

# Course map


Part 0: review
Part I: optimization tools
1 Calculus
2 Linear algebra
3 MLE
4 GD + ridge
5 Newton
6 Prox
7 SGD + Adam
Part II: constraints
Part III: uncertainty
8 LP + duality
9 Cones
10 Integer
11 Mean + risk
12 SAA
13 Context
Part IV: decision-focused learning
14 Predict, optimize
15 Learn the policy
16 Repair layers
17 Solution map
18 Weighted logistic

# Jump to the minimum of the parabola




fit a parabola, move to its minimum, repeat

# The Newton step

parabola: $q(w) = f(w_t) + f'(w_t)(w - w_t) + \frac{1}{2} f''(w_t)(w - w_t)^2$

its minimum: $q'(w) = f'(w_t) + f''(w_t)(w - w_t) = 0$

$$w_{t+1} = w_t - \frac{f'(w_t)}{f''(w_t)}$$

like gradient descent with $\eta = 1/f''(w_t)$

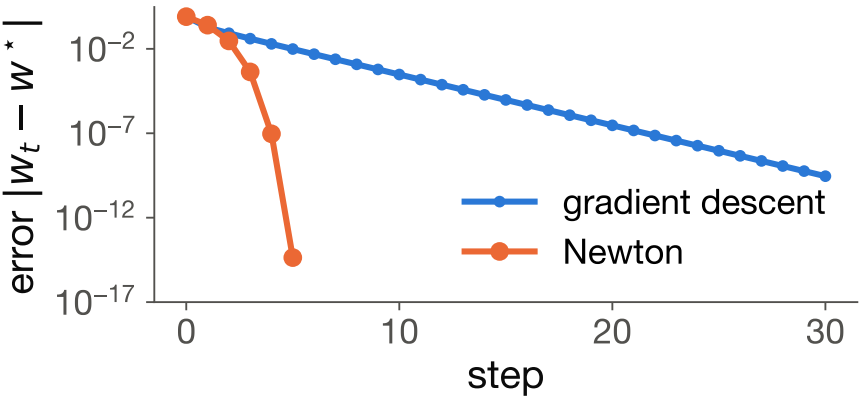
# A worked example

$$f(w) = e^w - 2w, \quad \text{minimizer } w^\star = \log 2$$

$$f'(w) = e^w - 2, \quad f''(w) = e^w$$

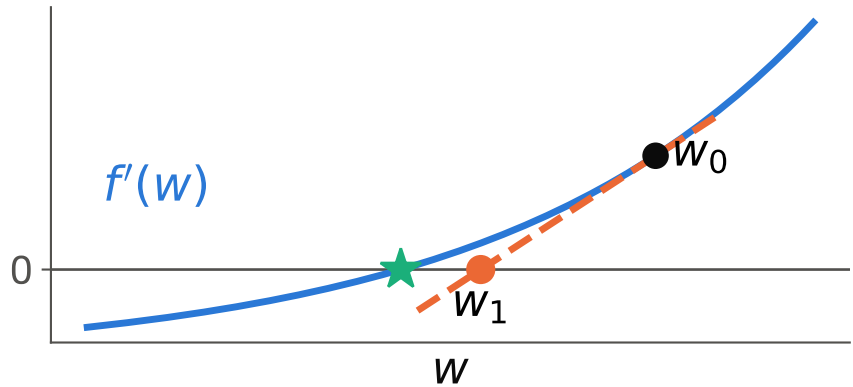
| $t$ | $w_t$ | error $\lvert w_t - w^\star \rvert$ |
|---|---|---|
| 0 | 1.5 | $8 \times 10^{-1}$ |
| 1 | 0.946 | $3 \times 10^{-1}$ |
| 2 | 0.7226 | $3 \times 10^{-2}$ |
| 3 | 0.69358 | $4 \times 10^{-4}$ |
| 4 | 0.6931473 | $9 \times 10^{-8}$ |
| 5 | 0.693147180560 | $4 \times 10^{-15}$ |

# Quadratic convergence: digits double



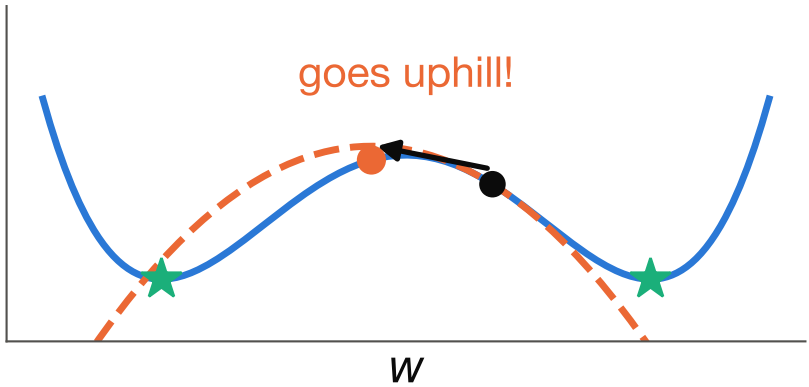

$\text{error}_{t+1} \approx c \cdot \text{error}_t^2$ $\quad 10^{-4} \to 10^{-8} \to 10^{-15}$

# Same idea: root finding on $f'$




Newton–Raphson: follow the tangent of $f'$ to zero

# When Newton goes wrong

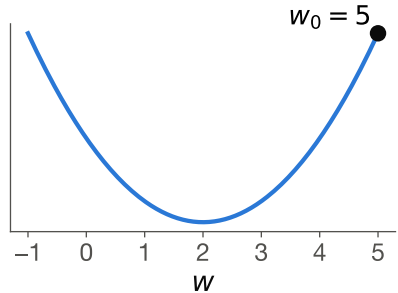



$f'' < 0$: the step goes toward a maximum

# Summary: Newton in one variable

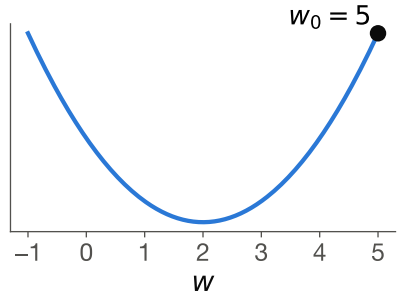
- Let $f(w)$ be a function, $w_t$ our point after $t$ steps.
- Let $q(w)$ be the parabola with the value, slope and curvature of $f$ at $w_t$.
- We want: $w^\star$, the minimizer of $f$, where $f'(w^\star) = 0$.
- Newton: $w_{t+1} = w_t - f'(w_t)/f''(w_t)$, the minimizer of $q$.
- Near $w^\star$ the convergence is quadratic; where $f'' < 0$, the step goes toward a maximum.

# Checkpoint 1




$f(w) = w^2 - 4w, \;\; w_0 = 5.$ One Newton step: $w_1 = ?$

# Checkpoint 1




$f(w) = w^2 - 4w, \ \ w_0 = 5.$ One Newton step: $w_1 = ?$

$w_1 = 5 - 6/2 = 2$: the exact minimizer

# Newton in many variables

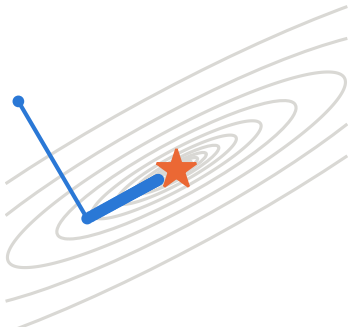
gradient descent: $w \leftarrow w - \eta \nabla f(w)$

Newton: $w \leftarrow w - H^{-1} \nabla f(w)$

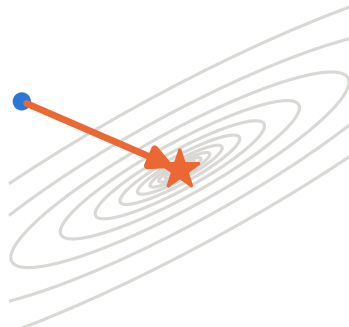
$H$ = Hessian (Lecture 2) · $H^{-1}$ replaces the single number $\eta$

# Newton on a quadratic: one step

gradient descent: 327 steps



Newton: 1 step



positive definite: one Newton step for any $\kappa$

# Why: $H^{-1}$ rescales each direction

what we see



what Newton sees



$H^{-1}$: each direction divided by its curvature

# The cost: each step is more work

| | gradient descent | Newton |
|---|---|---|
| each step | one gradient | solve a $p \times p$ system |
| number of steps | many (grows with $\kappa$) | few |
| a million weights | fine | too slow |

# Summary: Newton in many variables

- Let $w$ have $p$ entries, $\nabla f(w)$ be the gradient, $H$ the $p \times p$ Hessian.
- We want: $\min_w f(w)$.
- Newton: $w \leftarrow w - H^{-1}\nabla f(w)$; GD has one number $\eta$ in place of $H^{-1}$.
- Positive definite quadratic: $H^{-1}$ makes $\kappa = 1$, one step.
- Each step solves a $p \times p$ system: slow for $10^6$ weights.

# Checkpoint 2

$$H = \begin{pmatrix} 2 & 0 \\ 0 & 50 \end{pmatrix} \qquad \nabla f = \begin{pmatrix} 4 \\ 100 \end{pmatrix}$$

Newton step $H^{-1}\nabla f = ?$

# Checkpoint 2

$$H = \begin{pmatrix} 2 & 0 \\ 0 & 50 \end{pmatrix} \qquad \nabla f = \begin{pmatrix} 4 \\ 100 \end{pmatrix}$$

Newton step $H^{-1}\nabla f = ?$

$(4/2,\ 100/50) = (2,\ 2)$: equal steps

# Newton for logistic regression

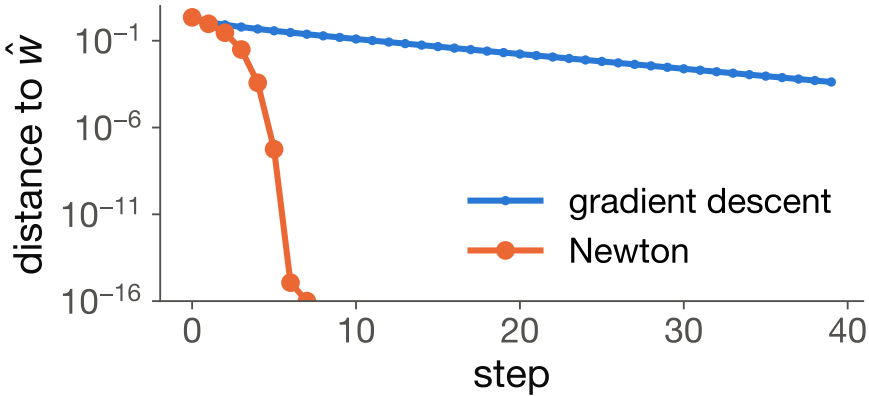
gradient (Lecture 3): $X^\top(p - y)$

Hessian: $X^\top S X, \quad S = \mathrm{diag}\big(p_i(1 - p_i)\big)$

$$w \leftarrow w - \big(X^\top S X\big)^{-1} X^\top (p - y)$$

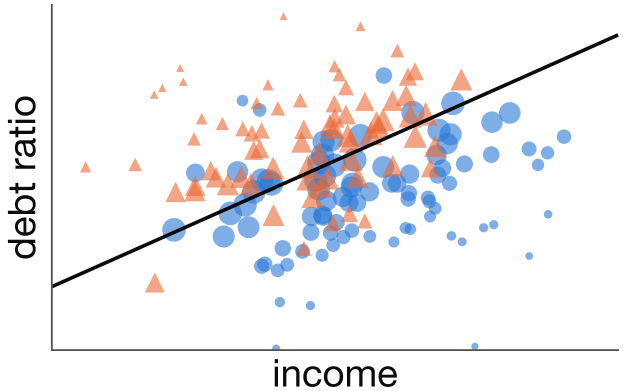
for logistic regression, Newton = Fisher scoring

# Newton vs GD on the credit data




Newton: $10^{-10}$ in 6 steps GD: $3 \times 10^{-9}$ after 100

# IRLS: weighted least squares




IRLS: iteratively reweighted least squares

weight $p_i(1-p_i)$: uncertain customers count most

# Summary: logistic regression and IRLS

- Let $x_i$ be row $i$ of $X$, $y_i = 1$ if customer $i$ defaulted.
- Let $p_i$ be the probability of default, $S = \mathrm{diag}\big(p_i(1-p_i)\big)$.
- We want: $\hat{w}$, the minimizer of the logistic loss $-\ell(w)$.
- Newton = Fisher scoring: $w \leftarrow w - \big(X^\top S X\big)^{-1} X^\top (p - y)$.
- IRLS: the same step as least squares, weights $p_i(1-p_i)$.

# Standard error and the 95% interval

new data would give a slightly different $\hat{\theta}$

SD (standard deviation): typical distance from the average

SE (standard error) of $\hat{\theta}$: its SD over many new data sets

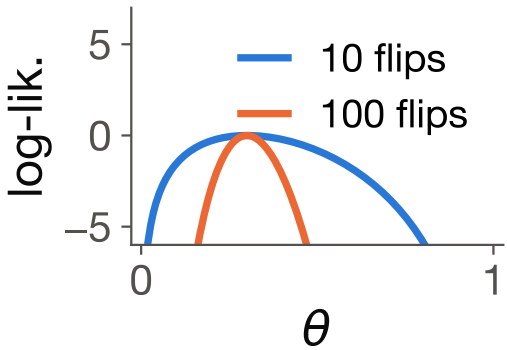
normal shape: 95% of the time within 1.96 SE of the truth

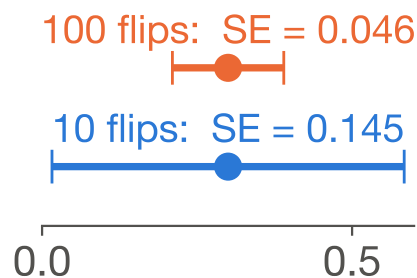
95% interval: $\hat{\theta} \pm 1.96\,\mathrm{SE}$

coin, 3 heads in 10: SE $= \sqrt{0.3 \cdot 0.7/10} = 0.145$

100 flips: SE $= 0.046$, $\sqrt{10}$ times smaller (next slide)

# Curvature = information

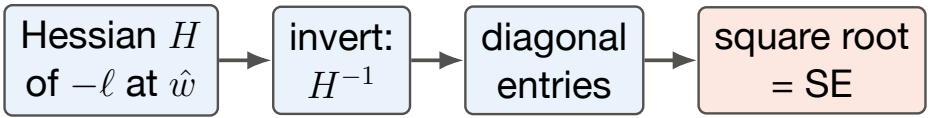



sharper peak $\Rightarrow$ smaller $\mathrm{SE} = 1/\sqrt{-\ell''(\hat{\theta})}$

right: 95% intervals, $\hat{\theta} \pm 1.96\,\mathrm{SE}$

# Standard errors from Newton's Hessian




| credit data | estimate | SE |
|---|---|---|
| intercept | $-0.51$ | 0.22 |
| income | $-1.27$ | 0.27 |
| debt ratio | $1.75$ | 0.30 |

income: $-1.27 \pm 1.96 \cdot 0.27$, from $-1.80$ to $-0.74$: below 0

# Variance of a misspecified model

bread: $H^{-1}$

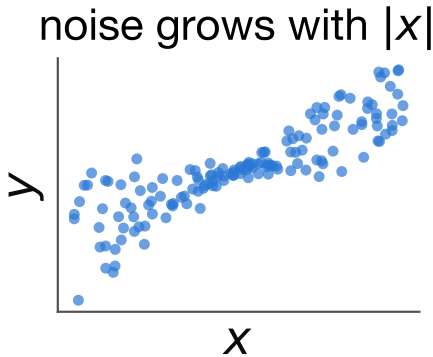
meat: $\sum_i g_i g_i^\top$, $g_i$: row $i$'s gradient at $\hat{w}$

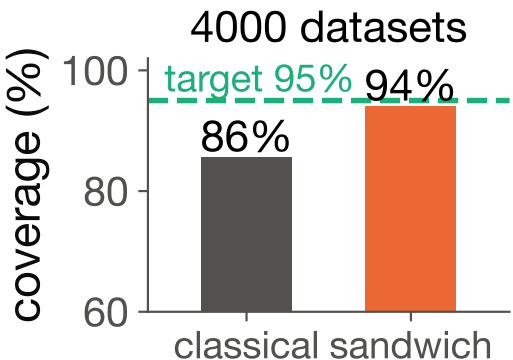
bread: $H^{-1}$

the sandwich: $\mathrm{Var}(\hat{w}) \approx H^{-1}\left(\sum_i g_i g_i^\top\right) H^{-1}$

model right: sandwich $\approx H^{-1}$

# Sandwich standard errors: a simulation



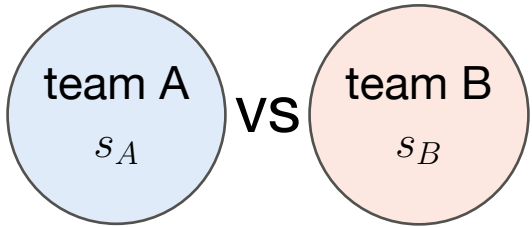

a straight line fit by least squares; the noise grows with $|x|$

coverage of 95% intervals: classical 86%, sandwich 94%

# Summary: standard errors

- Let $-\ell$ be the loss, $\hat{w}$ its minimizer, $H$ its Hessian at $\hat{w}$.
- We want: $\mathrm{Var}(\hat{w})$, how much $\hat{w}$ varies over new datasets.
- Classical: $\mathrm{Var}(\hat{w}) \approx H^{-1}$; each standard error is the square root of a diagonal entry.
- Sandwich: $H^{-1}\left(\sum_i g_i g_i^\top\right) H^{-1}$, $g_i$ = row $i$'s gradient at $\hat{w}$; any sum of row losses.
- Misspecified (noise SD grows with $|x|$): 95% intervals covered 86% (classical), 94% (sandwich).

# Ratings from wins: Bradley–Terry

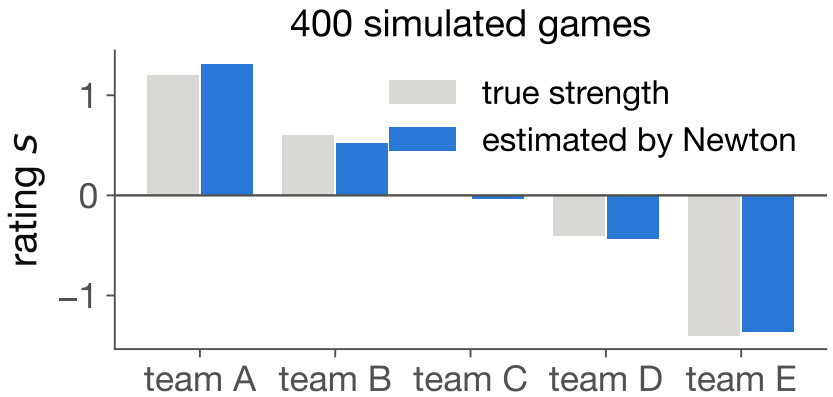



$$P(\text{A wins}) = \sigma\left(s_A - s_B\right)$$ ($\sigma$: the sigmoid)

logistic regression, no intercept: $y = 1$ if A won,
$x = (+1$ for A, $-1$ for B, $0$ else$)$

chess (Elo) and sports leagues

# Estimated versus true ratings




only differences matter: fix one team, fit, center

# Summary: the Bradley–Terry model

- Let $s_A$, $s_B$ be the ratings of teams A, B; $\sigma(z) = 1/(1+e^{-z})$.
- Model: $P(\text{A wins}) = \sigma(s_A - s_B)$: only differences matter.
- We want: the ratings that make past results most likely.
- Each game is a row: $+1$ for A, $-1$ for B, $0$ else.
- Each row sums to 0, so $H$ has no inverse: fix one team at 0, fit by Newton, then center.

# Checkpoint 3

$$s_A - s_B = 1$$

$P(\text{A wins}) = ?$ $\quad (e \approx 2.72)$

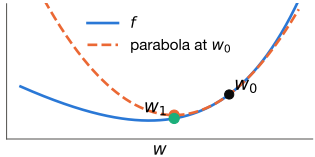
# Checkpoint 3

$$s_A - s_B = 1$$

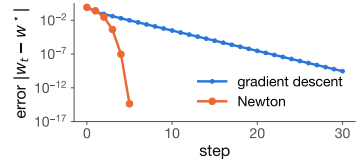
$P(\text{A wins}) = ?$ $\quad (e \approx 2.72)$

$$\sigma(1) = \frac{1}{1 + e^{-1}} \approx 0.73$$

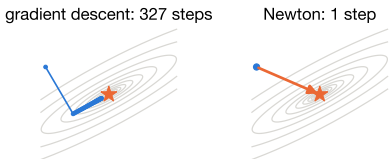
# Lecture summary



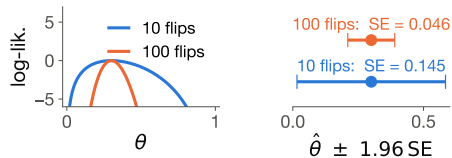

parabola's minimum




quadratic convergence (near $w^\star$)

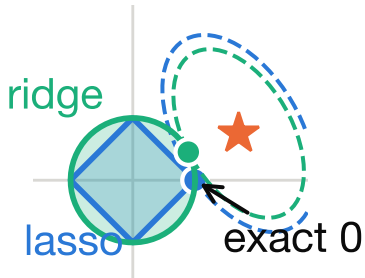


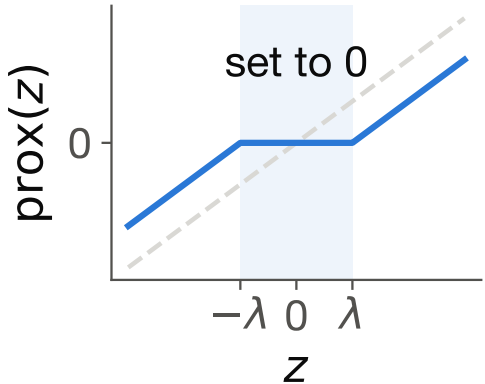

$w \leftarrow w - H^{-1}\nabla f$: few steps




curvature gives SEs

# Next lecture: non-differentiable losses




the lasso: no derivative at the kink, so a new tool

# Exercise

1. $f(w) = w^2 - 6w + 1,\ w_0 = 10$: one Newton step.
2. $f(w) = e^w - 3w$: write the Newton update; one step from $w_0 = 1$.
3. 40 heads in 100 flips: $\hat{\theta}$ and its standard error.
4. $s_A = 0.5,\ s_B = -0.5$: $P(\text{A wins})$?