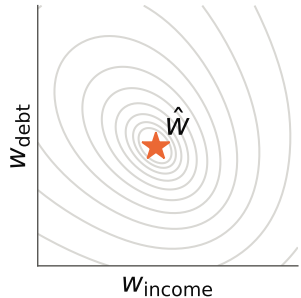
Gradient descent and ridge

208424 Lecture 4

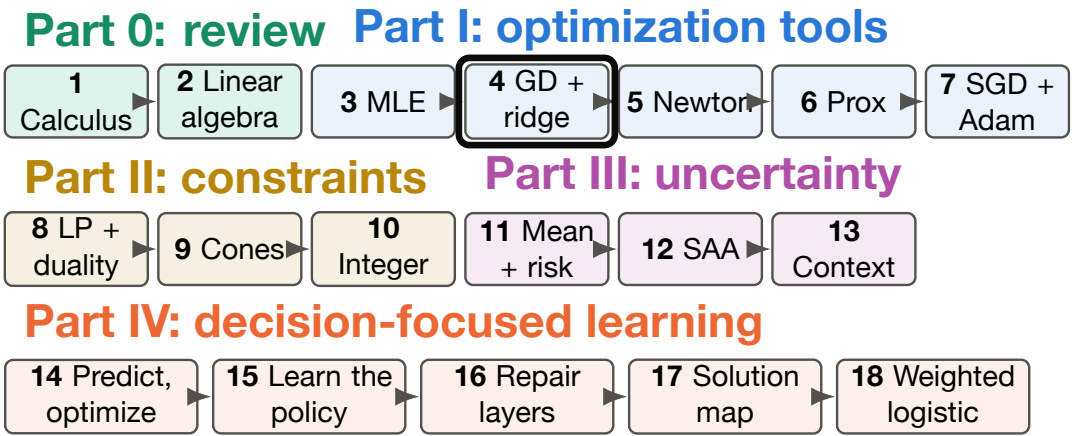
Lecture 3: no closed-form minimizer



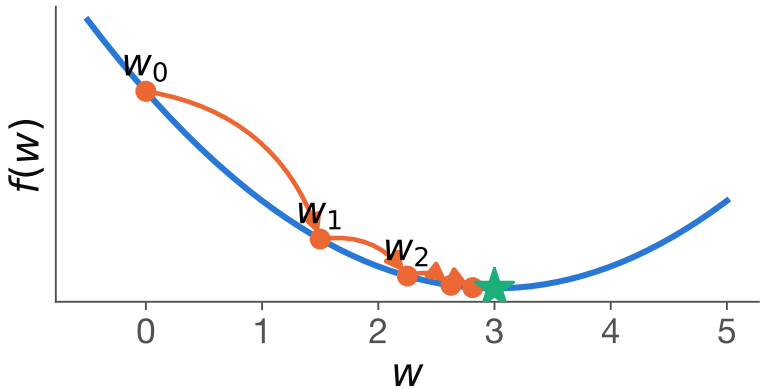
Lecture 3: logistic regression minimizes the loss $-\ell(w)$.
No closed form gives $\hat{w}$, the star, where the gradient is 0.

today: an algorithm that finds $\hat{w}$ step by step

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

Walk downhill, one step at a time



slope < 0: go right slope > 0: go left

The update rule

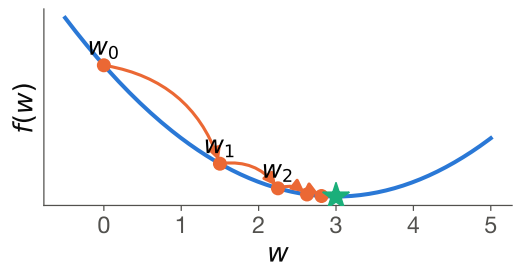
$$\underbrace{w_{t+1}}_{\text{new}} = \underbrace{w_t}_{\text{old}} - \underbrace{\eta}_{\text{step size}} \underbrace{\nabla f(w_t)}_{\text{uphill}}$$

gradient descent: the minus sign moves *against* the gradient

A worked example

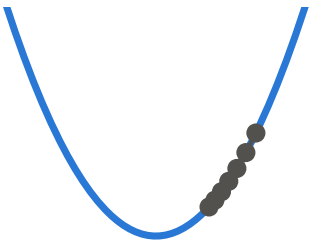
$$f(w) = (w-3)^2, \quad f'(w) = 2(w-3), \quad \eta = 0.25, \quad w_0 = 0$$

t	w_t	$f'(w_t)$	w_{t+1}
0	0	-6	1.5
1	1.5	-3	2.25
2	2.25	-1.5	2.625
			$\to 3$

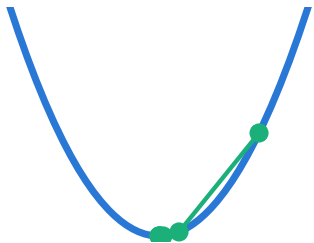


Three step sizes

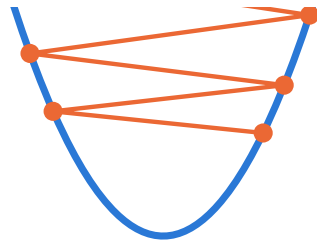
$\eta = 0.05$
too small



$\eta = 0.4$
good

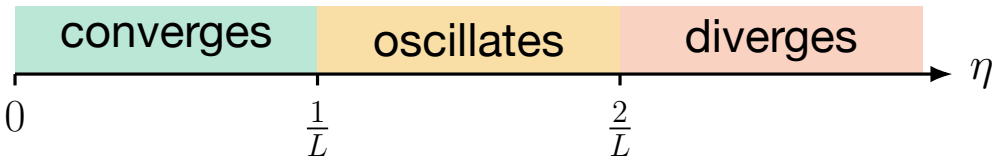


$\eta = 1.05$
too big



slope flips sign and grows $\Rightarrow$ step too big

How big a step?



L = largest curvature (largest Hessian eigenvalue)

exact on a quadratic · a good guide elsewhere

Why these step sizes

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

$$w_{t+1} = w_t - \eta \cdot 2w_t = (1 - 2\eta)\, w_t$$

$\eta = 0.25 < 1/L$	factor 0.5:	halves each step
$1/L < \eta = 0.75 < 2/L$	factor -0.5:	flips, shrinks: oscillates
$\eta = 1.05 > 2/L$	factor -1.1:	flips, grows: diverges

many weights: each eigenvector direction gets factor $1 - \eta\lambda$

$\eta = 1/L$: steepest direction 0, flattest $1 - 1/\kappa$ (slide 15)

Summary: gradient descent

- Let $f(w)$ be the loss, w_t our position after t steps.
- Let η be the step size; the gradient $\nabla f(w_t)$ points uphill.
- We want: $\min_w f(w)$, with no closed-form solution.
- Gradient descent: $w_{t+1} = w_t - \eta \nabla f(w_t)$.
- $L =$ largest curvature; on a quadratic, $\eta \leq 1/L$ converges and $\eta > 2/L$ diverges.

Checkpoint 1

$$f(w) = w^2, \quad w_0 = 4, \quad \eta = 0.25$$

$$w_1 = ? \qquad w_2 = ?$$

Checkpoint 1

$$f(w) = w^2, \quad w_0 = 4, \quad \eta = 0.25$$

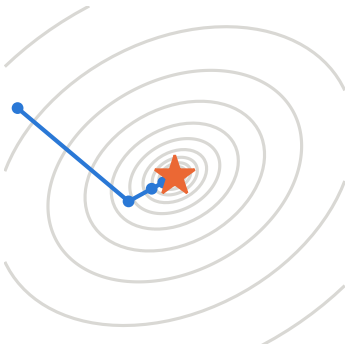
$$w_1 = ? \qquad w_2 = ?$$

$$f'(w) = 2w$$

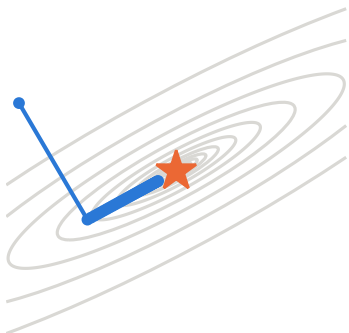
$$w_1 = 4 - 0.25 \cdot 8 = 2 \qquad w_2 = 2 - 0.25 \cdot 4 = 1$$

Well-conditioned versus ill-conditioned

$\kappa = 2$: 20 steps

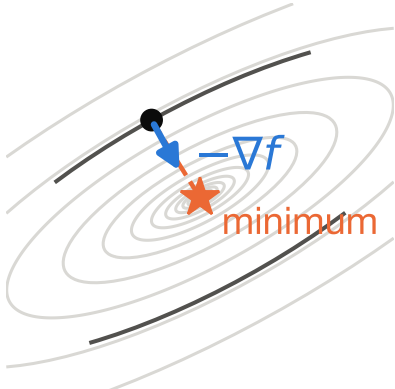


$\kappa = 25$: 327 steps



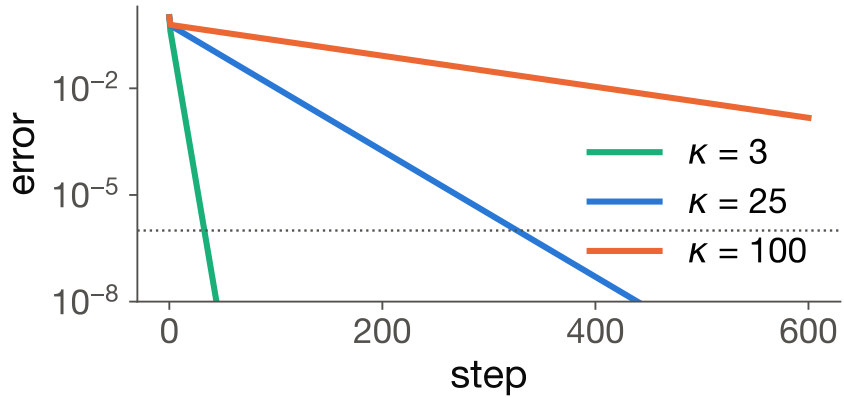
same accuracy: 20 steps versus 327

Why ill-conditioning is slow



$-\nabla f$ points across the valley, not at the minimum

Speed depends on κ

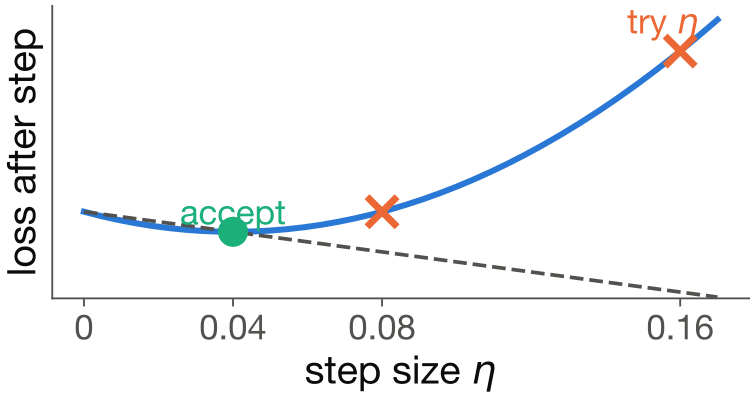


error $\times(1 - 1/\kappa)$ per step: 33, 327, 1327 steps

Summary: ill-conditioning is slow

- Let L be the largest Hessian eigenvalue, $\kappa = L \div$ the smallest, the condition number (Lecture 2).
- We want: the number of steps, for small κ (well-conditioned) and large κ (ill-conditioned).
- Ill-conditioned: $-\nabla f$ (downhill on the loss f) points across the valley, not at the minimum.
- Quadratic, step $\eta = 1/L$: error $\times (1 - 1/\kappa)$ per step.
- Steps grow like κ: 33, 327, 1327 steps for $\kappa = 3,\ 25,\ 100$.

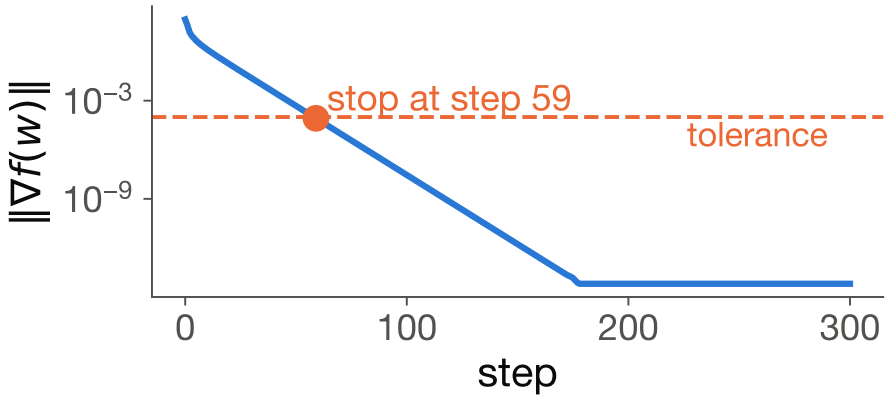
Backtracking: try a step, halve if needed



loss drops enough? accept · otherwise $\eta \leftarrow \eta/2$

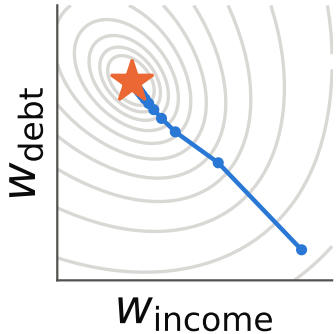
enough: $f(w - \eta g) \leq f(w) - \frac{1}{2}\eta \|g\|^2, \quad g = \nabla f(w)$

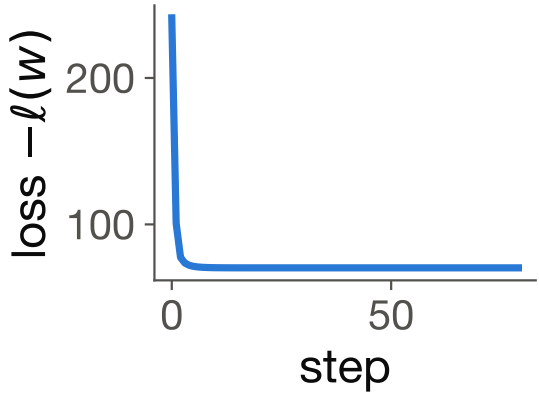
When to stop



stop when $\|\nabla f(w)\|$ is tiny: the loss is nearly flat

Gradient descent on the credit data





logistic loss: $L = \frac{1}{4}\,\lambda_{\max}(X^\top X),\ \ \eta = 1/L = 0.0214$

the gradient is below 10^{-4} after 59 steps

The whole algorithm

```
w = np.zeros(X.shape[1])
for t in range(1000):
    g = X.T @ (sigmoid(X @ w) - y)
    if np.linalg.norm(g) < 1e-4:
        break
    w = w - eta * g
```

eta $= 1/L$ $\cdot$ the gradient $X^\top(p - y)$ from Lecture 3

Summary: gradient descent in practice

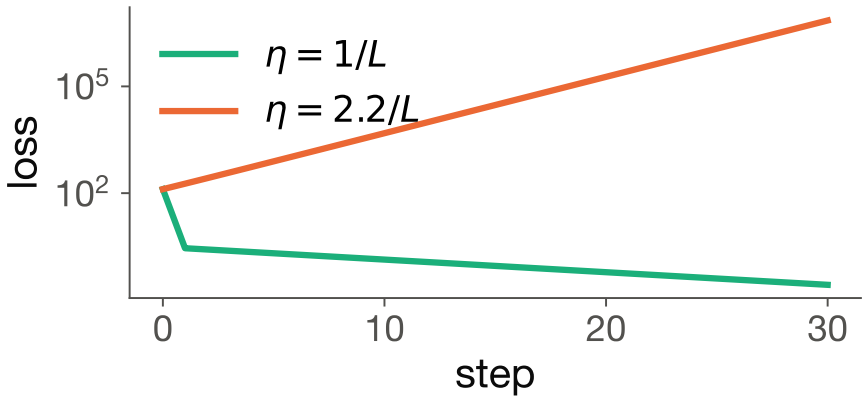
- Let $f(w) = -\ell(w)$, the logistic loss at the weights w.
- We want: $\hat{w}$, the minimizer of f, where $\nabla f(\hat{w}) = 0$.
- Repeat $w \leftarrow w - \eta \nabla f(w)$, $\eta = 1/L$, $L =$ largest curvature.
- No L? Backtracking: halve η until f drops enough.
- Stop when $\|\nabla f(w)\| < 10^{-4}$: 59 steps on the credit data.

Checkpoint 2

Your loss gets *larger* every step.

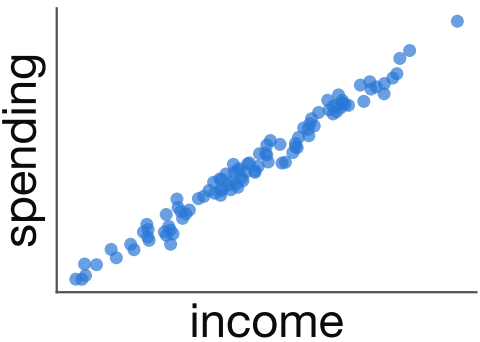
What went wrong?

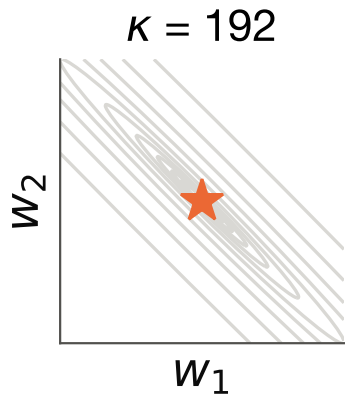
Checkpoint 2: the step is too big



η above $2/L$: halve η (or use backtracking)

Correlated features: ill-conditioning

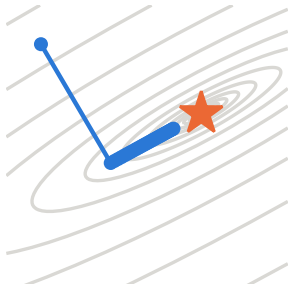




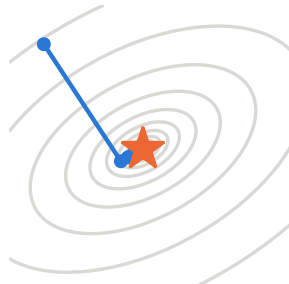
income and spending: correlation 0.99 $\Rightarrow$ $\kappa = 192$

Ridge speeds up gradient descent

least squares
$\kappa = 25$: 327 steps



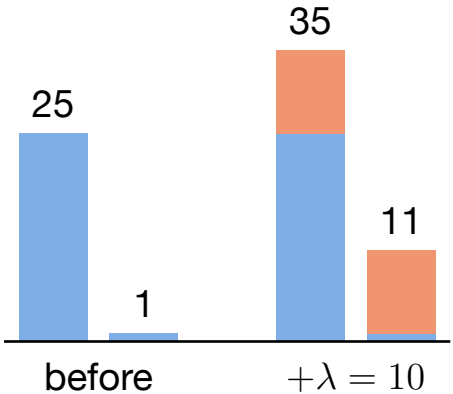
ridge
$\kappa = 3.2$: 33 steps



minimize $f(w) + \lambda \|w\|^2$

327 steps $\rightarrow$ 33 steps

Why: ridge adds λ to every eigenvalue



$$\kappa = \frac{\sigma_{\max}^2 + \lambda}{\sigma_{\min}^2 + \lambda}$$

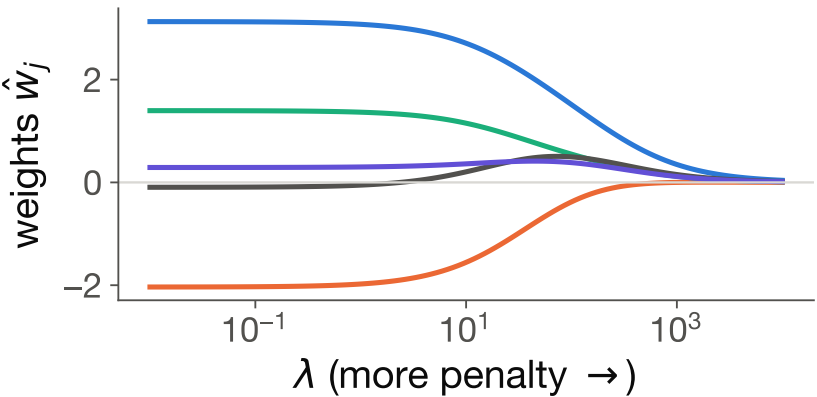
$$\frac{25}{1} = 25 \longrightarrow \frac{35}{11} = 3.2$$

$$X^\top X v = \sigma^2 v \Rightarrow$$

$$(X^\top X + \lambda I)\, v = (\sigma^2 + \lambda)\, v$$

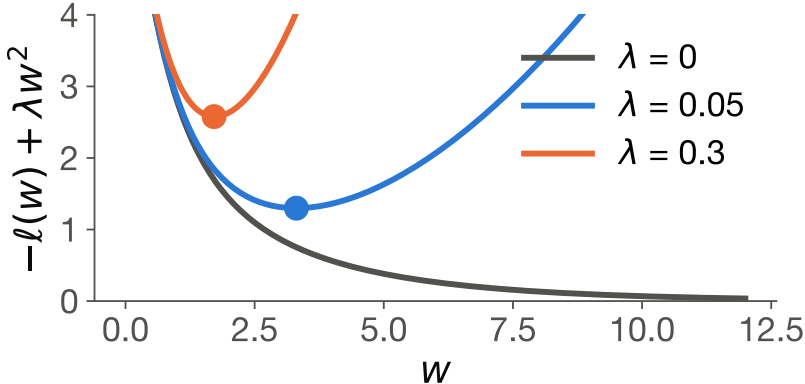
σ^2 = eigenvalues of $X^\top X$ · the 2 in $2X^\top X$ and $2\lambda I$ cancels

Ridge shrinks the weights



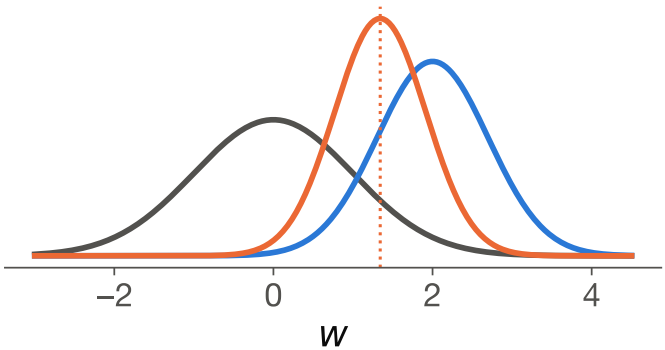
larger $\lambda \Rightarrow$ shorter $\|w\|$; all weights $\to 0$

Ridge fixes separable data (Lecture 3)



$\lambda > 0$: the penalized loss has a minimizer

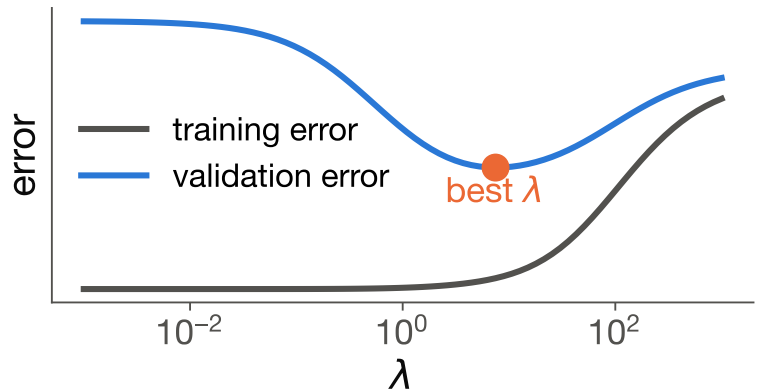
Ridge as a prior belief



prior: small weights
likelihood: data
posterior

$f = -\ell$, prior $w \sim N(0, \tau^2) \Rightarrow$ ridge with $\lambda = \dfrac{1}{2\tau^2}$

Choosing λ: check on new data



pick the λ with the smallest *validation* error

Summary: ridge

- Let σ^2 be the eigenvalues of $X^\top X$, X the data, w the weights.
- We want: $\min_w \; f(w) + \lambda \|w\|^2$, f the loss, $\lambda > 0$: ridge.
- Each eigenvalue grows by λ: $\kappa = (\sigma^2_{\max} + \lambda)/(\sigma^2_{\min} + \lambda)$.
- Larger λ: smaller κ, shorter $\|w\|$; separable data get a minimizer.
- Too large λ: weights near 0, the data barely count; choose λ by validation error.

Checkpoint 3

eigenvalues of $X^\top X$: 100 and 1

$\kappa = ?$ with $\lambda = 9$: $\kappa = ?$

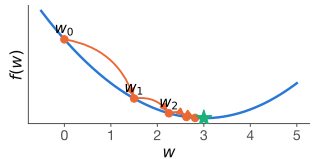
Checkpoint 3

eigenvalues of $X^\top X$: 100 and 1

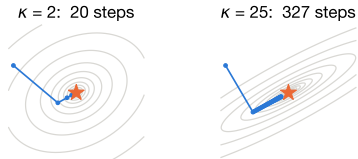
$\kappa = ?$ with $\lambda = 9$: $\kappa = ?$

$100/1 = 100$ $109/10 = 10.9$

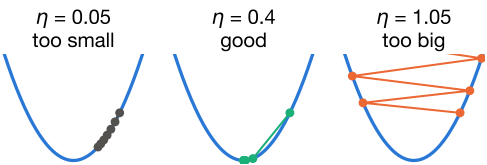
Lecture summary



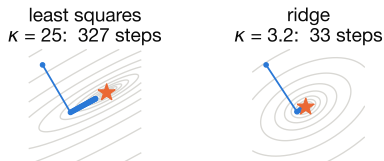
$$w \leftarrow w - \eta \nabla f(w)$$



ill-conditioned (κ large) = slow

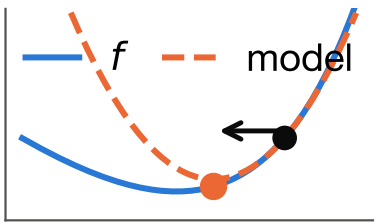


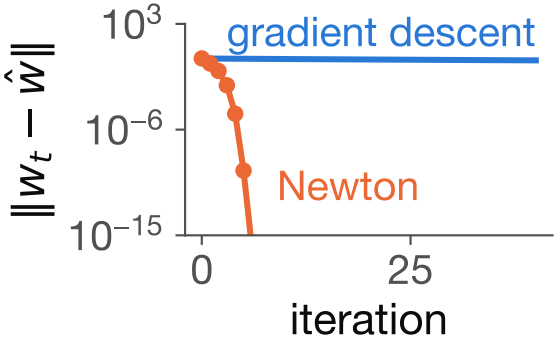
step size: $\eta \leq 1/L$



ridge lowers κ

Next lecture: Newton's method





use curvature: move to the parabola's minimum in one step

Exercise

1. $f(w) = (w+1)^2$, $w_0 = 1$, $\eta = 0.25$: find w_1 and w_2.
2. $f(w) = 5w^2$: what is L? Above which η does GD diverge?
3. Eigenvalues of $X^\top X$ are 50 and 2. Find κ, then κ with $\lambda = 6$.
4. Run the code of slide 20 on the Lecture 3 credit data ($w_0 = 0$, $\eta = 1/L$). How many steps?