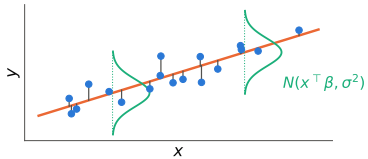
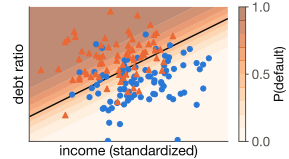
Estimation is optimization

208424 Lecture 3

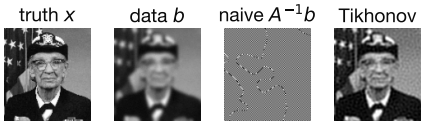
Every fitted model is an optimization



regression



credit risk



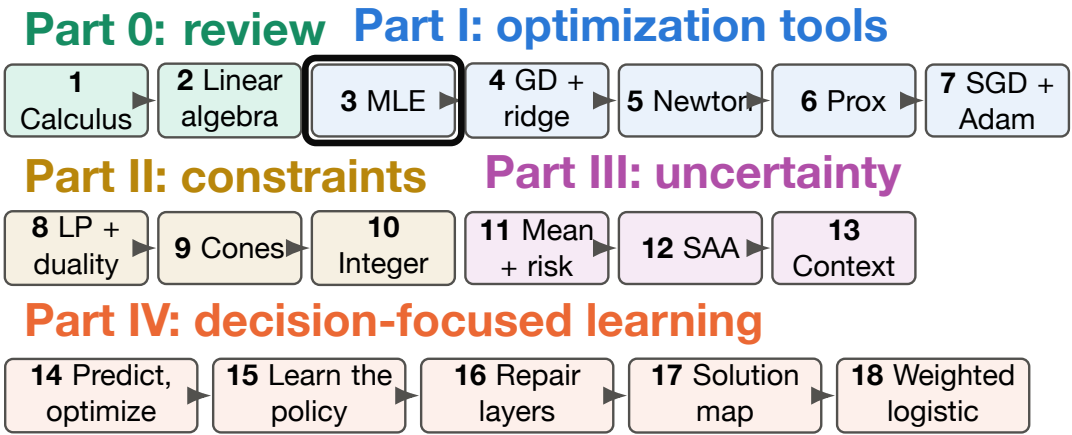
deblurring



chatbots

$$\hat{\theta} = \arg\min_\theta \ \text{loss}(\theta)$$

Course map

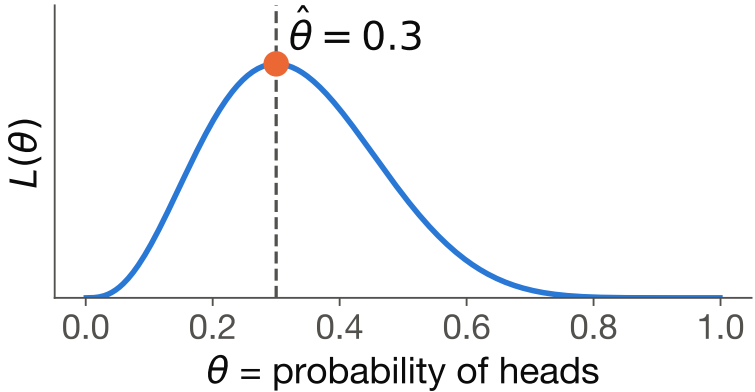

Part 0: review
Part I: optimization tools
1 Calculus
2 Linear algebra
3 MLE
4 GD + ridge
5 Newton
6 Prox
7 SGD + Adam
Part II: constraints
Part III: uncertainty
8 LP + duality
9 Cones
10 Integer
11 Mean + risk
12 SAA
13 Context
Part IV: decision-focused learning
14 Predict, optimize
15 Learn the policy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

Ten coin flips



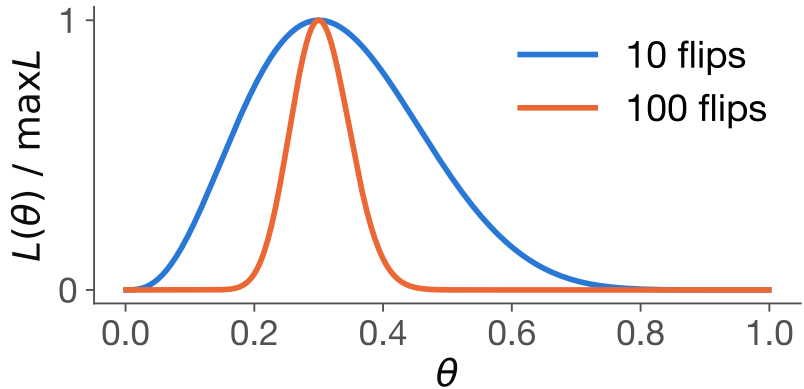
$$\theta = P(\text{heads}) = ?$$

Likelihood of θ



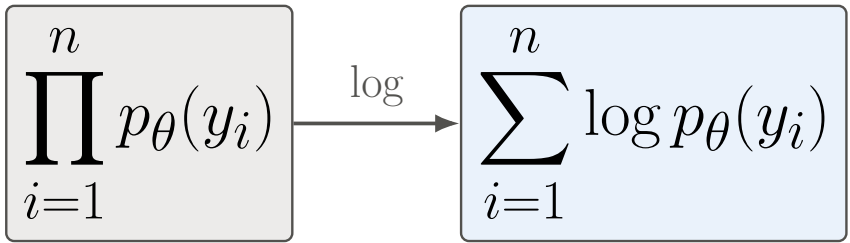
$$L(\theta) = \theta^3(1-\theta)^7$$

More data, sharper likelihood



same $\hat{\theta} = 0.3$, less uncertainty

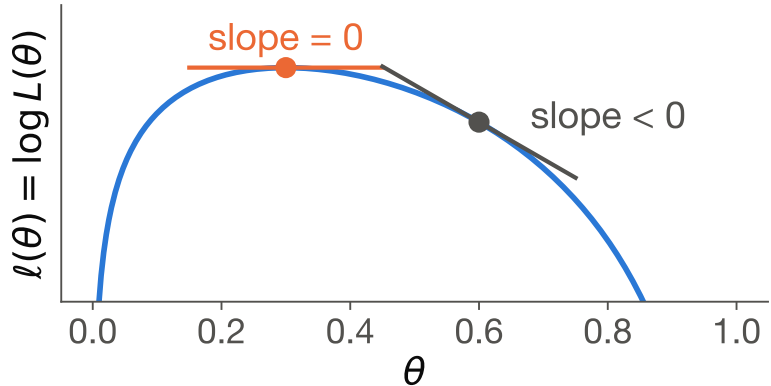
Take logs: products become sums



$0.5^{1100} = 0.0$ in float64 $\qquad 1100 \log 0.5 = -762.46$

same maximizer, no underflow, easier derivatives

The maximizer has slope zero



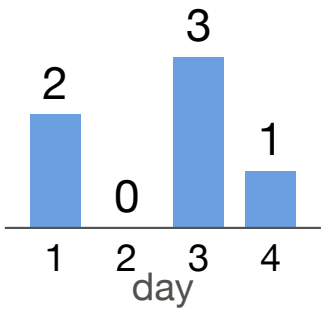
k heads in n flips: $\ell(\theta) = k \log \theta + (n-k) \log(1-\theta)$

$$\ell'(\theta) = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \implies \hat{\theta} = \frac{k}{n}$$

Summary: maximum likelihood

- Let $y_1, \ldots, y_n$ be independent data, $p_\theta(y)$ the model, and θ the unknown parameter.
- Likelihood: $L(\theta) = \prod_i p_\theta(y_i)$, how probable our data are.
- Log-likelihood: $\ell(\theta) = \sum_i \log p_\theta(y_i)$, same maximizer.
- We want: $\hat{\theta} = \arg\max_\theta \ \ell(\theta)$; at the maximum, $\ell'(\hat{\theta}) = 0$.
- Coins, k heads in n flips: $\hat{\theta} = k/n$, here $3/10 = 0.3$.

Checkpoint: accidents per day

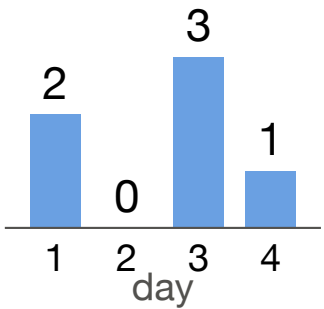


$$y_i \sim \text{Poisson}(\lambda)$$

$$p_\lambda(y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$$\hat{\lambda} = ?$$

Checkpoint: accidents per day



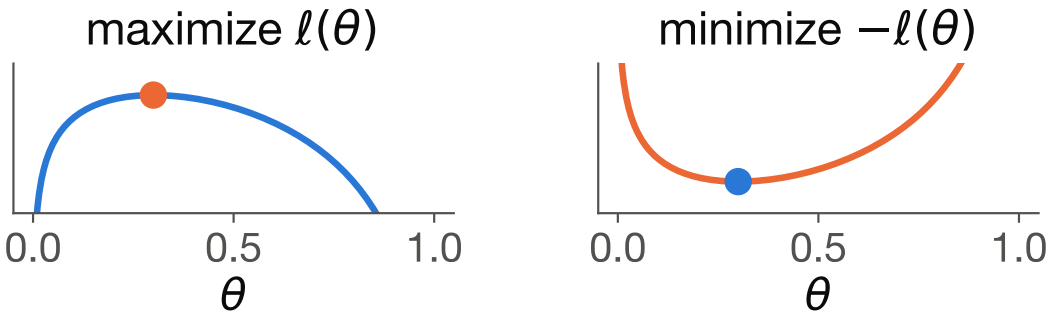
$$y_i \sim \text{Poisson}(\lambda)$$

$$p_\lambda(y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$$\hat{\lambda} = ?$$

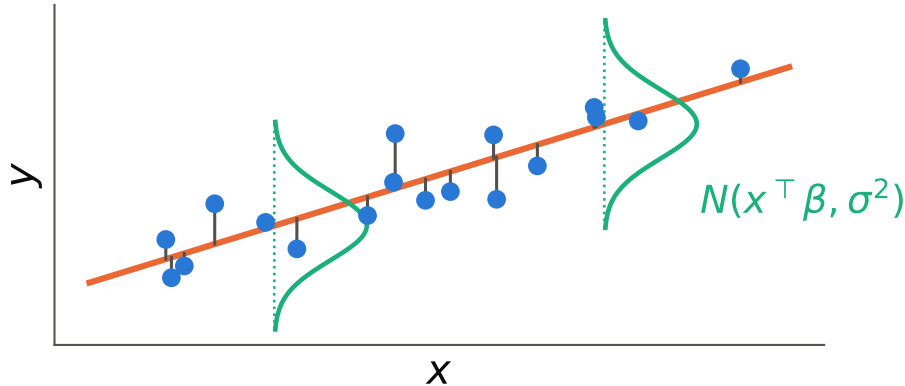
$$\ell(\lambda) = \sum_i \left(y_i \log \lambda - \lambda\right) + \text{const} \implies \hat{\lambda} = \bar{y} = \frac{6}{4} = 1.5$$

Maximizing ℓ = minimizing $-\ell$



software minimizes, so we minimize the *loss* $-\ell$

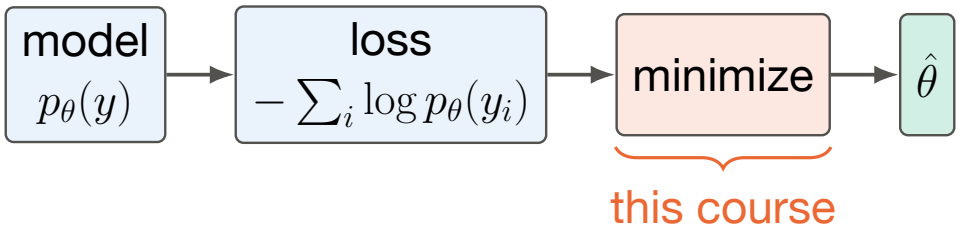
Normal noise: MLE is least squares



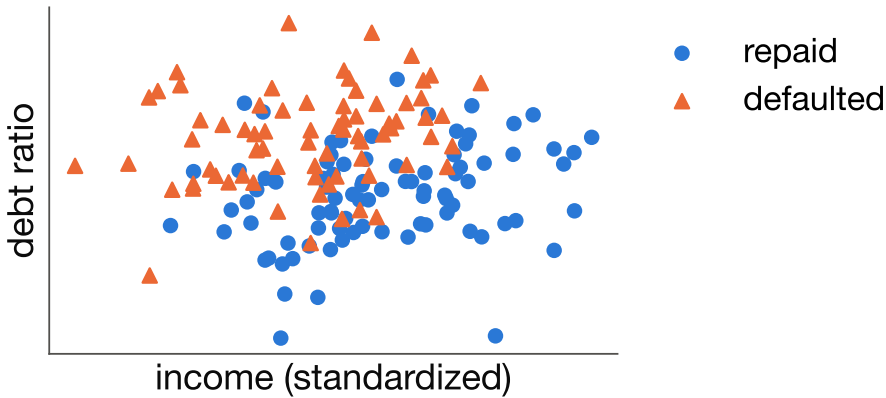
normal noise: $p_\beta(y_i) \propto e^{-(y_i - x_i^\top \beta)^2/(2\sigma^2)}$, σ: noise SD

$$-\ell(\beta) = \frac{1}{2\sigma^2} \sum_i \left(y_i - x_i^\top \beta\right)^2 + \text{const}$$

The MLE recipe

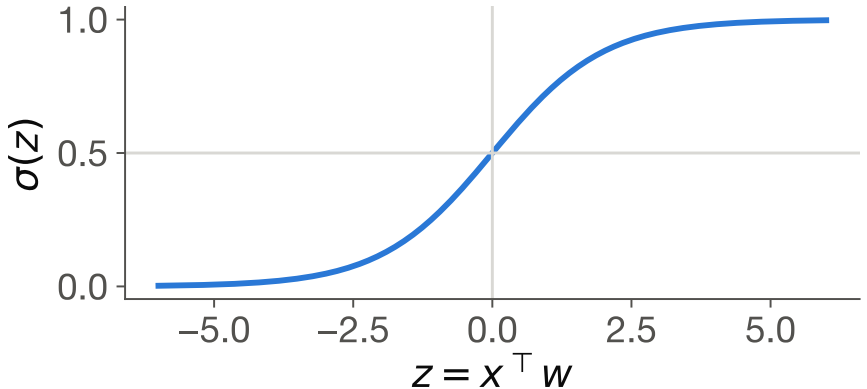


Who will default on a loan?



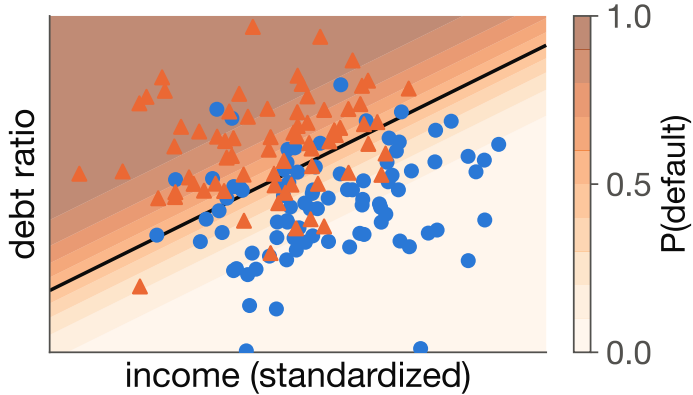
simulated data, 160 customers

Probabilities need values in $[0, 1]$



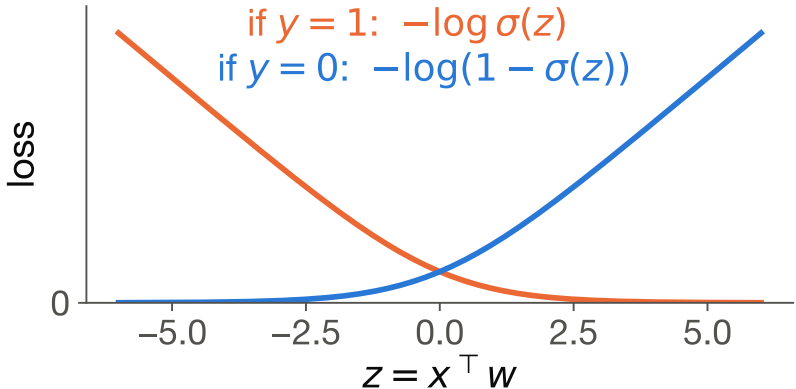
$\sigma(z) = \dfrac{1}{1 + e^{-z}}$ maps any z into $(0, 1)$

Logistic regression



$$P(y = 1 \mid x) = \sigma\left(x^\top w\right)$$

The loss of one customer



confident and wrong $\Rightarrow$ large loss

One formula for both outcomes

$$z = x^\top w, \quad p = \sigma(z) = \frac{e^z}{1+e^z}, \quad 1 - p = \frac{1}{1+e^z}$$

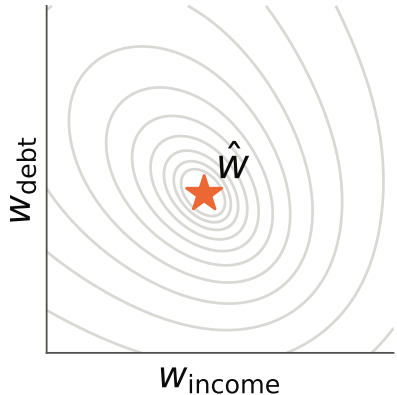
$$y = 1: \quad -\log p = \log\left(1 + e^z\right) - z$$

$$y = 0: \quad -\log(1-p) = \log\left(1 + e^z\right)$$

both: $\log\left(1 + e^z\right) - y\,z$

one customer's loss; the next slide adds them up

Total loss: no closed-form minimizer

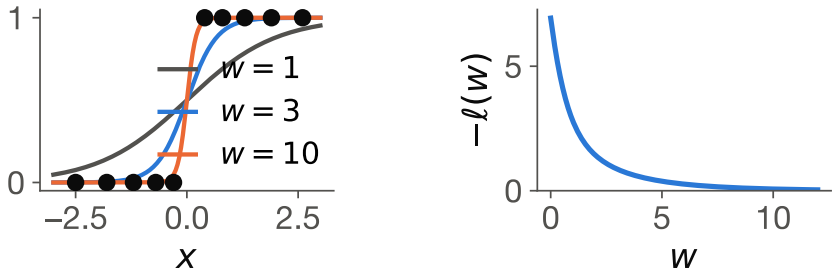


$$-\ell(w) = \sum_i \left[\log\left(1 + e^{x_i^\top w}\right) - y_i \, x_i^\top w \right]$$

Summary: logistic regression

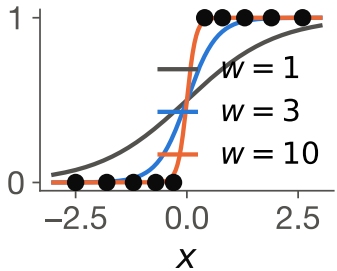
- Let x_i be customer i's features, $y_i = 1$ if i defaulted, else 0.
- Model with weights w: $P(y_i = 1 \mid x_i) = \sigma(x_i^\top w)$, $\sigma(z) = 1/(1 + e^{-z})$.
- Loss: $-\ell(w) = \sum_i \left[\log\left(1 + e^{x_i^\top w}\right) - y_i \, x_i^\top w \right]$.
- We want: $\hat{w} = \arg\min_w \left(-\ell(w)\right)$, the MLE of w.
- No closed form for $\hat{w}$: an algorithm must find it step by step.

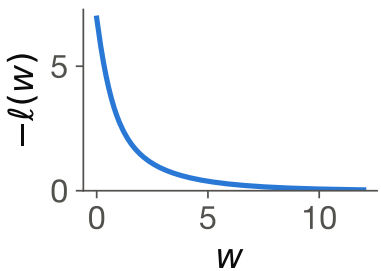
Checkpoint: separable data



What does the optimizer do?

Checkpoint: separable data

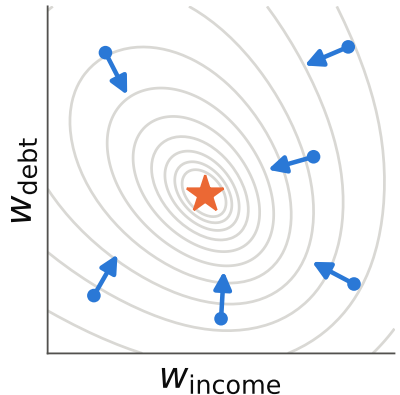




What does the optimizer do?

$-\ell(w) \to 0,\ \|w\| \to \infty$: no MLE. Fix: add $\lambda\|w\|^2$ (Lecture 4)

The negative gradient points downhill



$$\nabla\big(-\ell\big)(w) = \sum_i \big(p_i - y_i\big)\, x_i = X^\top (p - y) \qquad p_i = \sigma(x_i^\top w)$$

Deriving the gradient and the Hessian

one customer: $\log\left(1+e^z\right) - y\,z, \quad z = x^\top w$

slope in z: $\sigma(z) - y = p - y$

Lecture 1, slide 19: the slope of $\log(1+e^z)$ is $\sigma(z)$

chain rule, slope of z in w is x: $(p-y)\,x$

all customers: $\sum_i (p_i - y_i)\,x_i = X^\top(p-y)$, Lecture 2's $X^\top r$

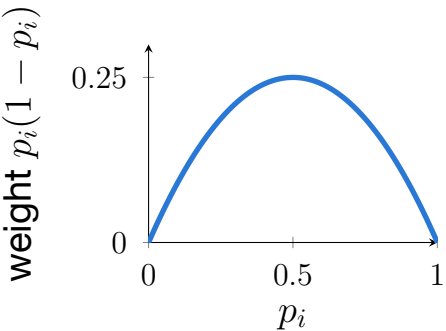
slope of $p = \sigma(z)$ in z: $e^z/(1+e^z)^2 = p\,(1-p)$

so the Hessian is $\sum_i p_i(1-p_i)\,x_i x_i^\top = X^\top S X$ (next slide)

Hessian: the curvature

$$\nabla^2\big(-\ell\big)(w) = X^\top S\, X$$

$$S = \mathrm{diag}\big(p_i(1-p_i)\big)$$

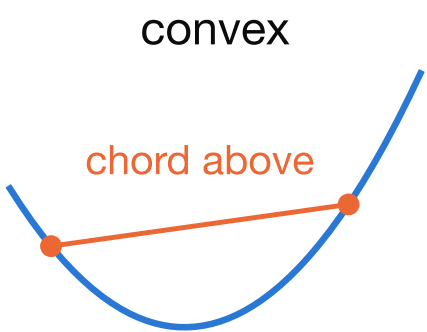


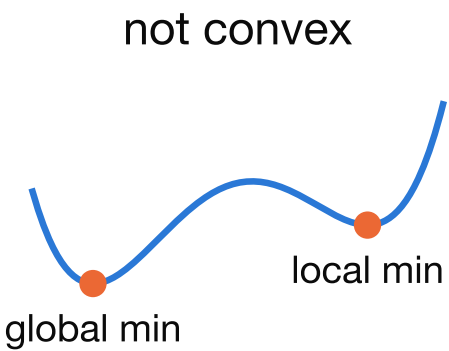
$$v^\top X^\top S X v$$

$$= \textstyle\sum_i p_i(1-p_i)\,(x_i^\top v)^2$$

≥ 0 for every v

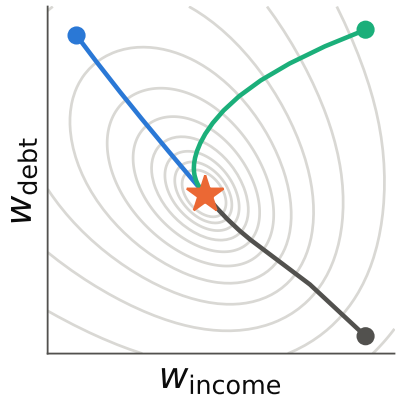
Convexity: every local min is global





Hessian $\succeq 0$ $\Rightarrow$ convex $\Rightarrow$ local min = global min

Why convexity matters

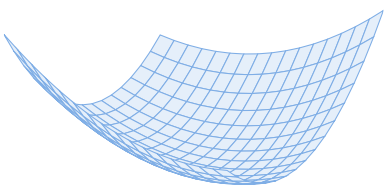


these data: every start reaches the same $\hat{w}$

Summary: gradient, Hessian, convexity

- Let X have rows $x_i^\top$; p, y have entries $p_i = \sigma(x_i^\top w)$, y_i.
- We want: every local min of $-\ell$ to be a global min.
- Gradient $X^\top(p - y)$ (uphill); Hessian $X^\top S\, X$, $S = \mathrm{diag}\big(p_i(1 - p_i)\big)$.
- For every v: $v^\top X^\top S X v = \sum_i p_i(1 - p_i)(x_i^\top v)^2 \geq 0$, so Hessian $\succeq 0$.
- So $-\ell$ is convex: any local min is a global min, if a minimum exists.

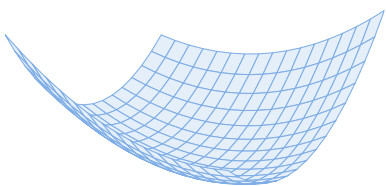
Checkpoint: is least squares convex?



$$f(\beta) = \|y - X\beta\|^2$$

$$\nabla^2 f = ?$$

Checkpoint: is least squares convex?



$$f(\beta) = \|y - X\beta\|^2$$

$$\nabla^2 f = ?$$

$$\nabla^2 f = 2X^\top X$$

$$v^\top X^\top X v = \|Xv\|^2 \geq 0$$

yes, convex

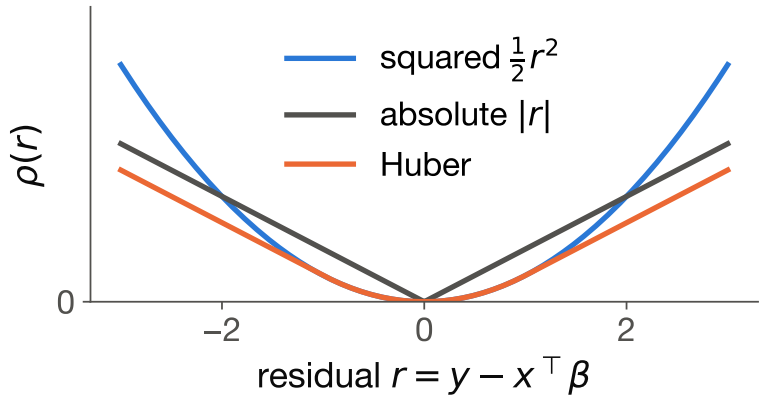
M-estimation: choose the loss

$$\hat{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \rho\big(y_i, x_i; \theta\big)$$

$-\log p_\theta$	r^2	$\lvert r \rvert$	Huber	log loss
MLE	least squares	median	robust	classification

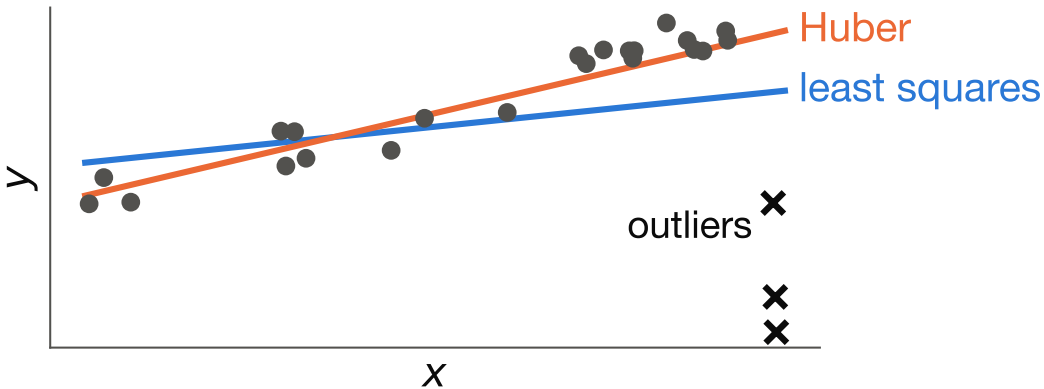
ρ = loss for one observation $(r = y - x^\top \beta)$

Three losses for regression



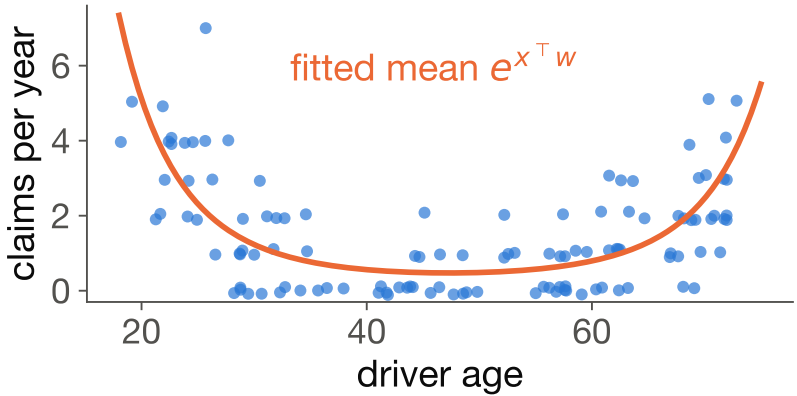
Huber: squared for small r, linear for large r

Three outliers: points far from the rest



true slope 0.5: least squares 0.19, Huber 0.44

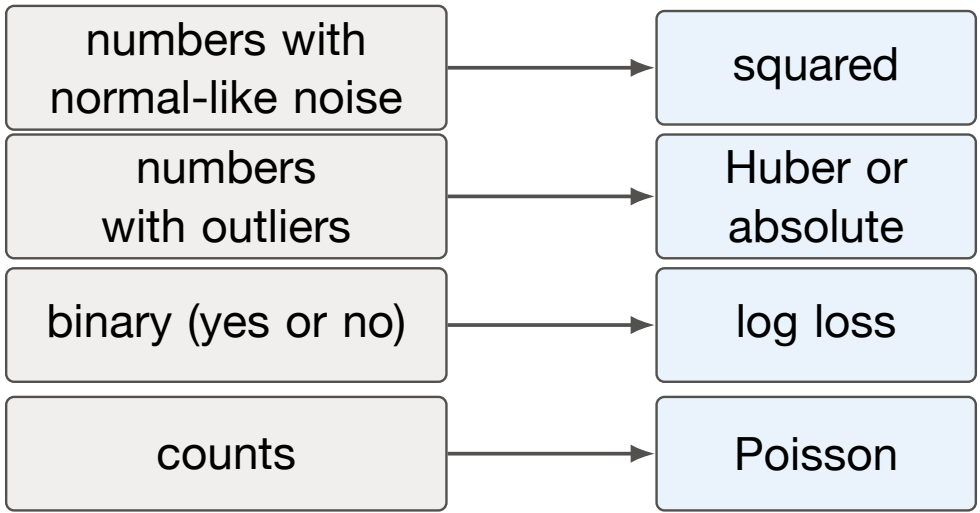
Counts: Poisson regression



$$-\ell(w) = \sum_i \left[e^{x_i^\top w} - y_i \, x_i^\top w \right] + \text{const}$$

simulated

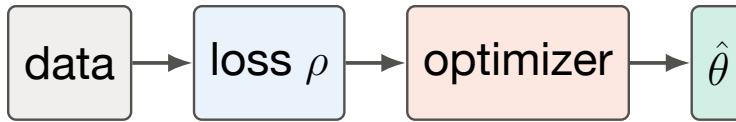
Which loss?


numbers with normal-like noise
squared
numbers with outliers
Huber or absolute
binary (yes or no)
log loss
counts
Poisson

Summary: M-estimation

- Let $\rho(y_i, x_i; \theta)$ be the loss of one observation.
- We want: $\hat{\theta} = \arg\min_\theta \ \frac{1}{n} \sum_{i=1}^n \rho(y_i, x_i; \theta)$, an M-estimator.
- $\rho = -\log p_\theta$ gives the MLE: log loss, Poisson loss.
- Regression ($\theta = \beta$, residual $r = y - x^\top \beta$): r^2 gives least squares, $|r|$ the median (for a constant), Huber is robust.
- With three outliers, least squares gives slope 0.19 and Huber 0.44; the true slope is 0.5.

Lecture summary



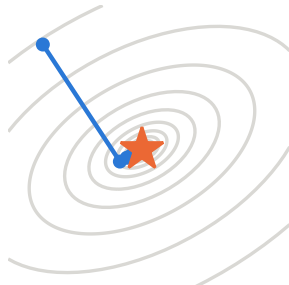
convex loss $\Rightarrow$ any local minimum is a global minimum

Next lecture: GD and conditioning

least squares
$\kappa = 25$: 327 steps



ridge
$\kappa = 3.2$: 33 steps



how many steps does gradient descent need?

Exercise

1. Derive $\nabla(-\ell) = X^\top(p - y)$ by hand.
2. Code it and check against finite differences:

```
def grad(w):
    p = 1 / (1 + np.exp(-X @ w))
    return X.T @ (p - y)
```