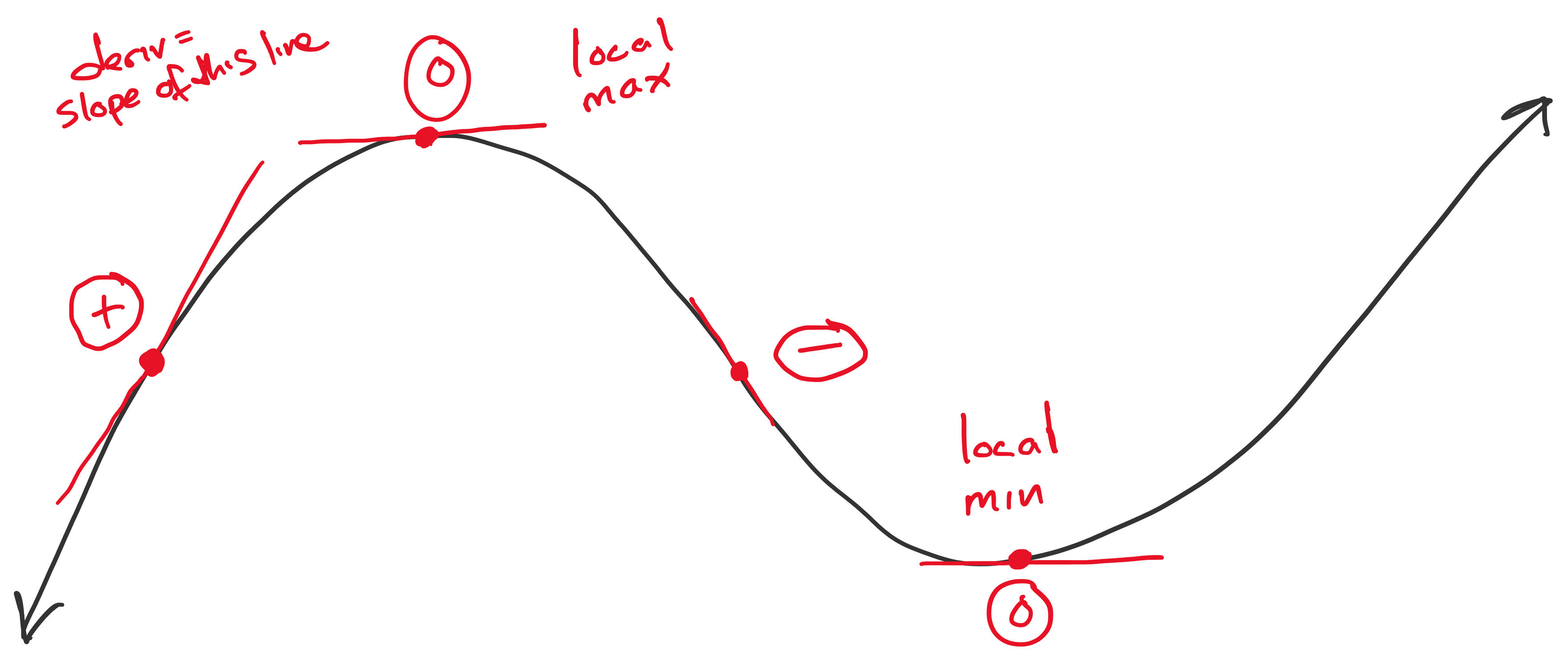
Grad Desc - WC

Thursday, October 17, 2019 9:56 AM

Defs:

Derivative: The instantaneous rate of change at a point on a function.



Optimize Either max or min

Loss Function Any function that describes the discrepancy btwn y and $\hat{y}$.

$$MSE = \frac{1}{n}\sum (y_i - \hat{y}_i)^2$$

$$\text{LASSO loss} = MSE + \lambda \|\beta\|_1$$

$$\text{Class. Error} = 1 - acc$$ ← Not differentiable!

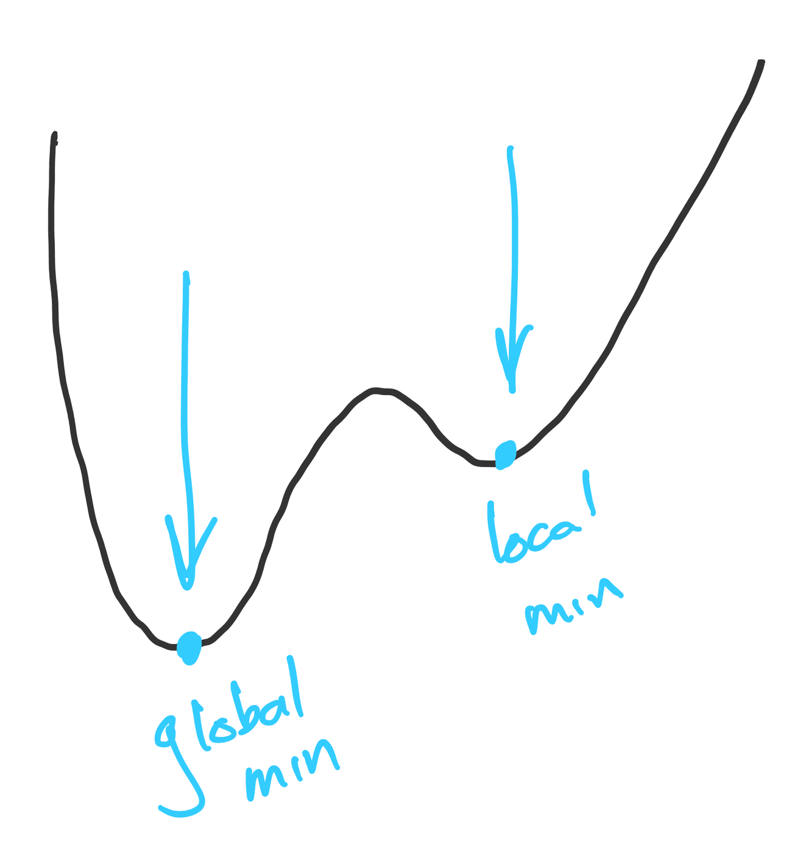
Log Loss (Binary Crossentropy)
$$= \frac{1}{n}\sum \left[-y_i \log \hat{p}_i - (1-y_i)\log(1-\hat{p}_i) \right]$$

$$\text{SVM Hinge Loss} = \frac{1}{n}\sum \max\{0, 1 - y_i f(x_i)\} + \frac{C}{2}\|w\|$$

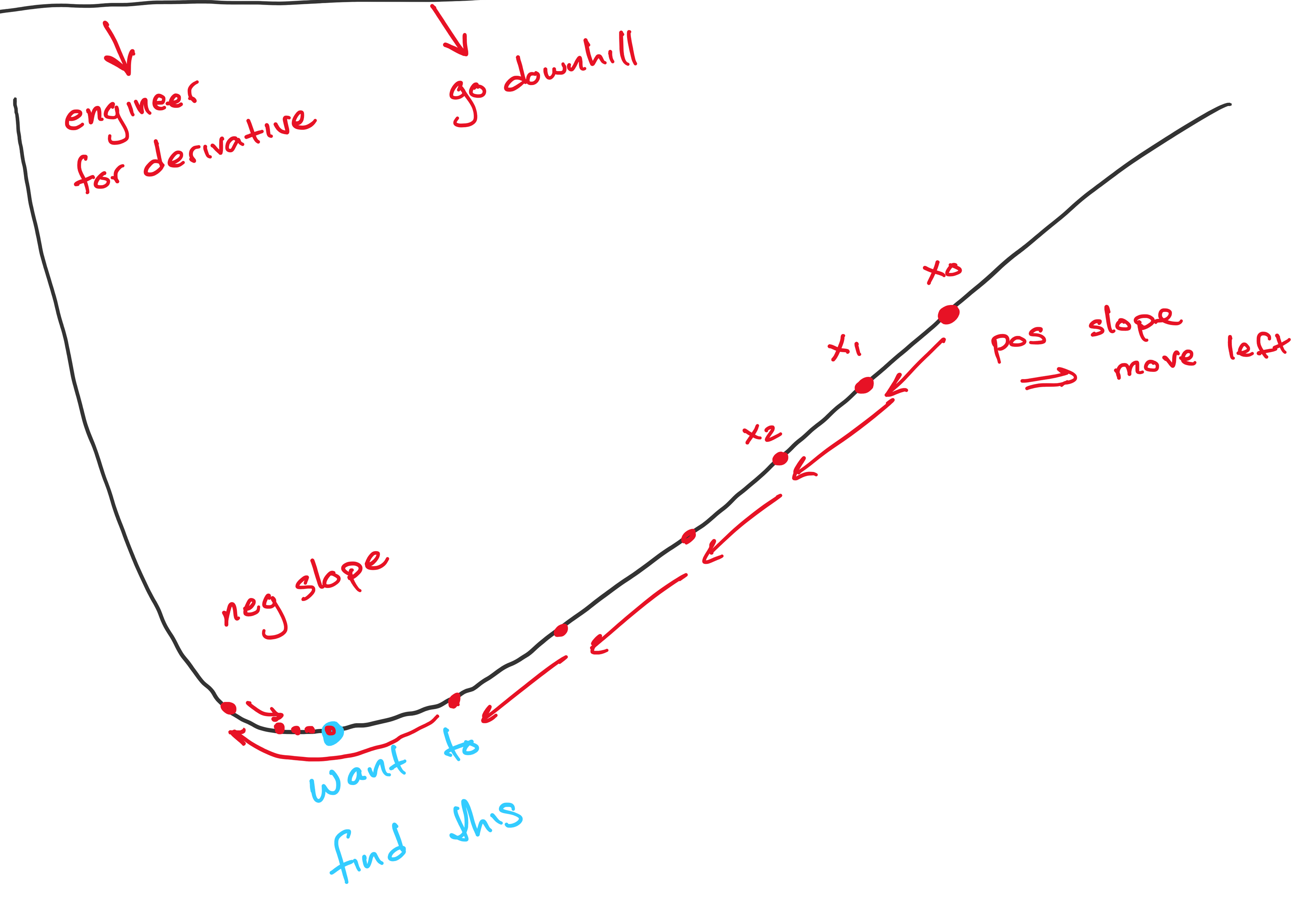
Convexity U-shaped function. Check $f''(x) \geq 0$ for all x

Convex

Nonconvex



Gradient Descent



Old way:

minimize f: $f'(x) = 0 \implies$ solve for x

works for MSE: $\hat{\beta} = (X^T X)^{-1} X^T y$

Don't believe me? $xe^x = 5$

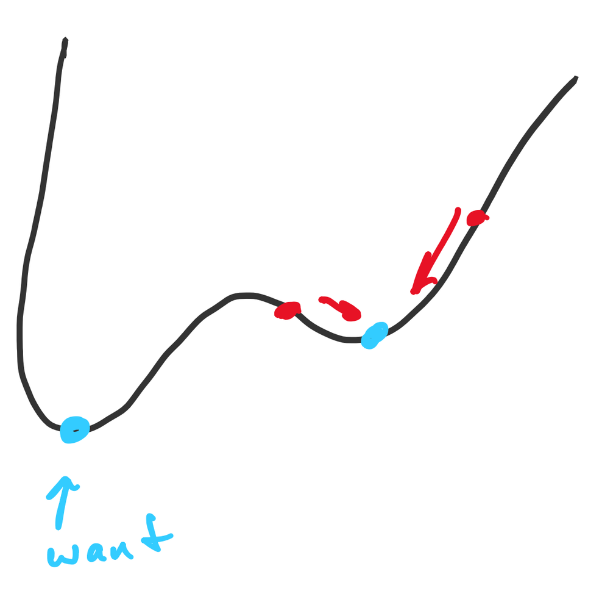
At step $k+1$:

$$x_{k+1} \leftarrow x_k - \alpha f'(x_k)$$

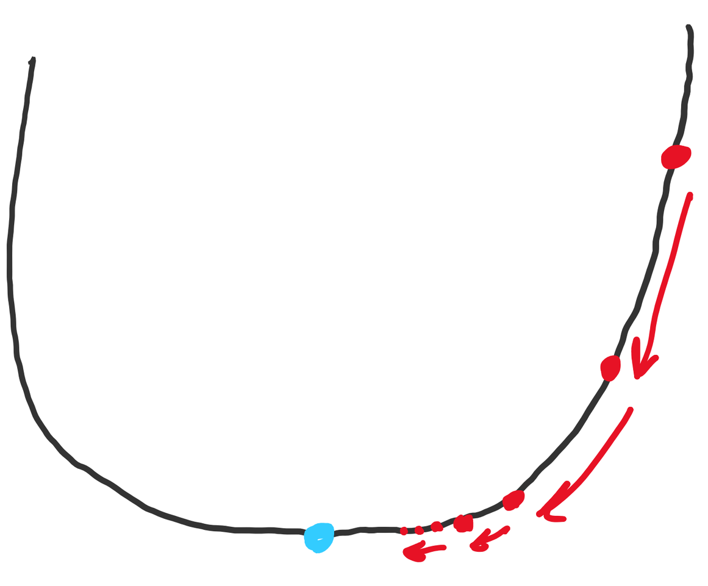
- x_{k+1}: new step
- x_k: current step
- $-$: opp dir
- α: step size
- $f'(x_k)$: magnitude and dir of change

What can go wrong?!

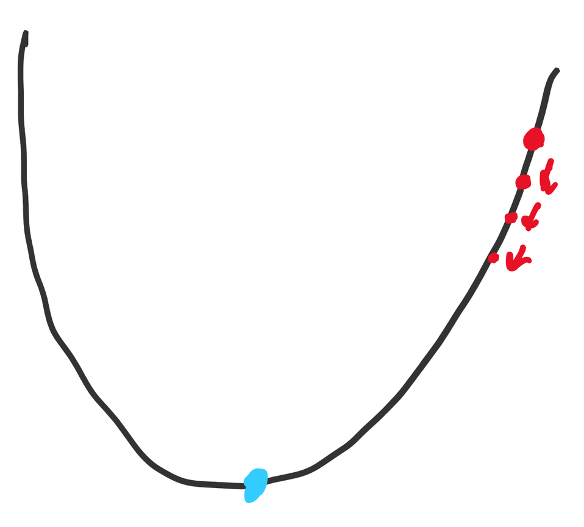
- Local minima



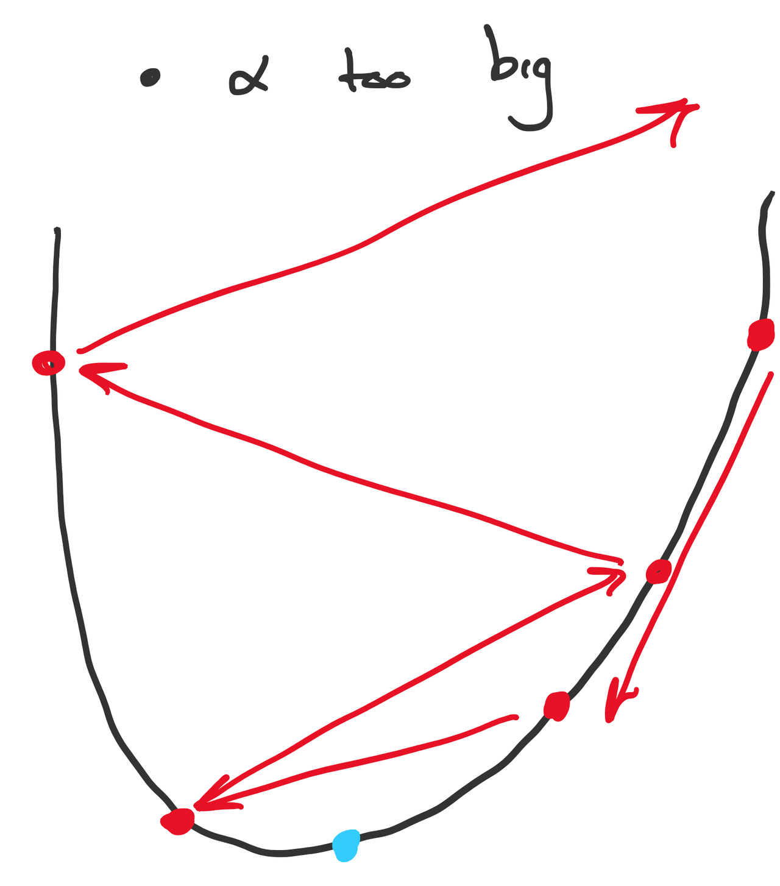
- Function too flat



- α too small



- α too big



How can we choose α?

- Guess and check
- $\alpha = \frac{c}{k}$ ← works sometimes; usually α shrinks too fast
- Newton's Method

$$x_{k+1} \leftarrow x_k - \frac{f'(x_k)}{f''(x_k)}$$ ← difficult or impossible to find

$$x_{k+1} \leftarrow x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

(x_{k+1}: vec; x_k: vec; $\nabla^2 f(x_k)$: matrix; $\nabla f(x_k)$: vec)

- Fancier algorithms
  - RMSProp
  - Adam

minimize f:

$$f(x) = -\frac{\log x}{1+x}$$

$$f'(x) = -\frac{1/x + 1 - \log x}{(1+x)^2}$$