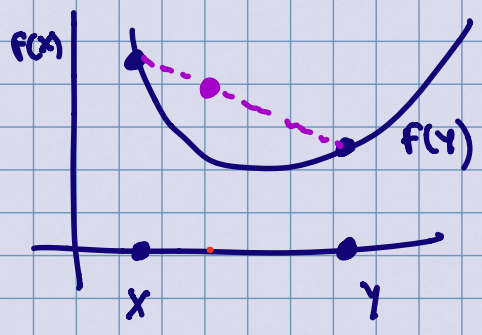
Lecture 29

Understanding Gradient Descent.

* MIDTERM ALREADY GRADED

* DROP DATE IS 11/18 (WEDNESDAY)

CONVEX FUNCTIONS



A function f is convex if

$$f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$$

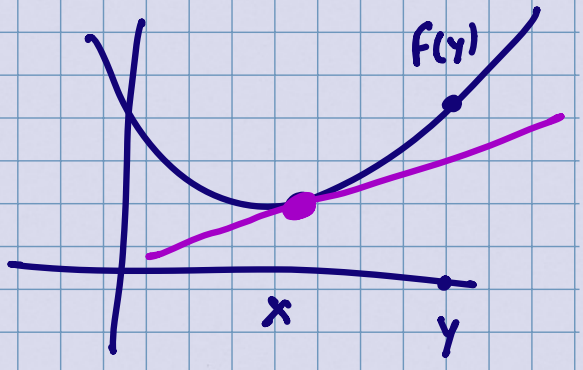
"function below the chord"

Characterizations (all equivalent, under suitable conditions)

- "first-order": for all x, y

$$f(y) \geq f(x) + \nabla f(x)^T (y-x)$$

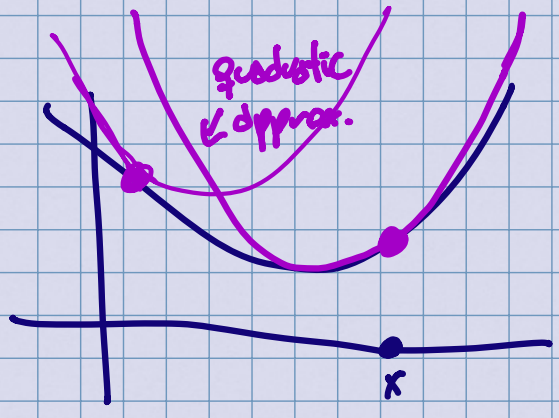
linear approximation at x



- "second order": for all x

$$\frac{\partial^2 f(x)}{\partial x^2} \text{ is a PSD matrix}$$

Hessian



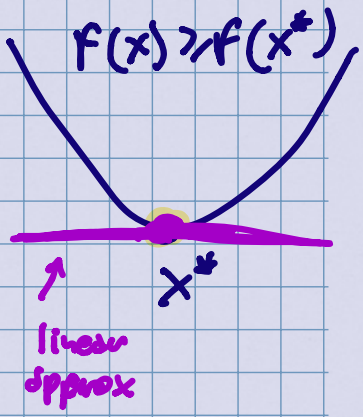
"quadratic approximation has positive curvature".

WHY IS CONVEXITY SO AWESOME?

- LOCAL MINIMA ARE ALSO GLOBAL

- SIMPLE CHARACTERIZATION OF OPTIMALITY

$$\nabla f(x^*) = 0$$



"gradient must vanish"
"zero gradient at the optimal point"

- A SIMPLE ALGORITHM ("gradient method")

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

next point — current point — stepsize — gradient at current point

EXAMPLE: SQUARE ROOTS.

HOW TO COMPUTE THE SQUARE ROOT OF A NUMBER?

E.G., $\sqrt{\beta}$? $\quad (\beta > 0)$

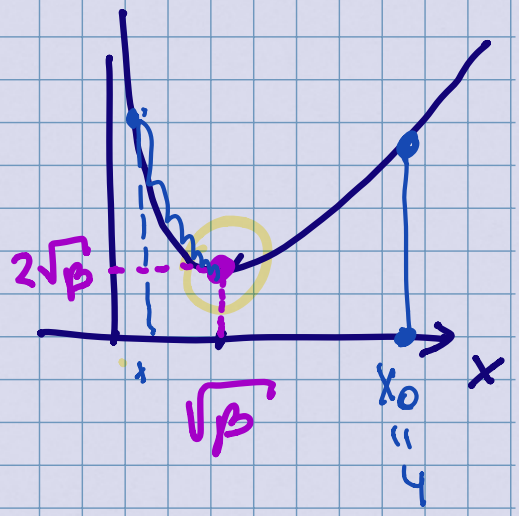
→ HERE'S AN APPROACH*, USING GRADIENT DESCENT

(* DISCLAIMER: BETTER METHODS EXIST!)

- STEP 1: GIVEN β, FIND A CONVEX FUNCTION WITH $\sqrt{\beta}$ AS THE MINIMUM (OR MINIMIZER)

E.G.: $(x > 0)$

$$F(x) = x + \frac{\beta}{x}$$



Why? $\quad DF(x) = 1 - \frac{\beta}{x^2} = 0 \quad \Rightarrow \quad x = \pm\sqrt{\beta}$

- **STEP 2:** LET'S WRITE GRADIENT DESCENT!

stepsize ↓

$$x_{k+1} = x_k - \gamma \, Df(x_k)$$

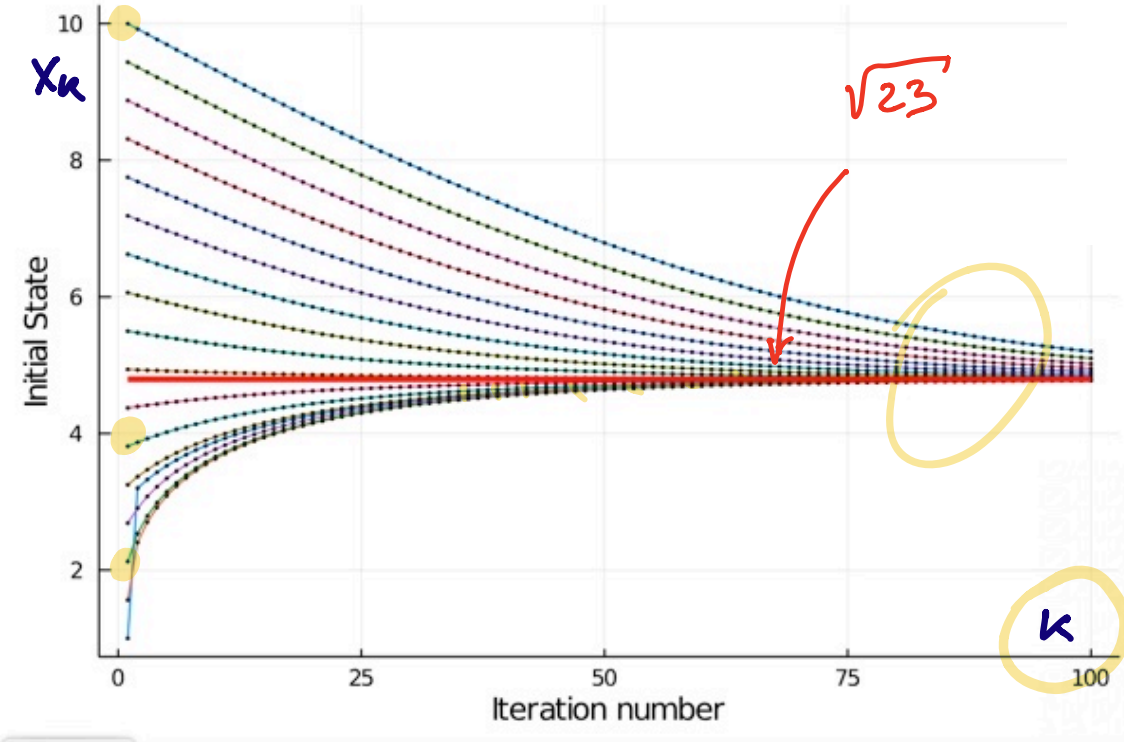
$$= x_k - \gamma \left(1 - \frac{\beta}{x_k^2}\right)$$

Take, e.g. $\beta = 23$

stepsize $\gamma = 0.1$

(from Julia)

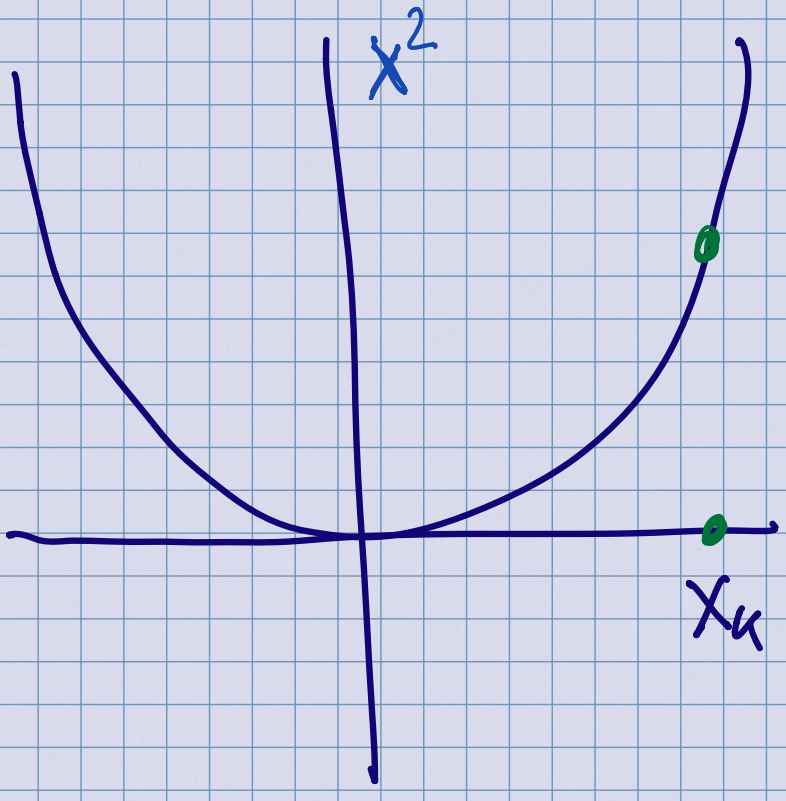
```
In [21]: trajectories[:,4]

Out[21]: 100-element Array{Float64,1}:
 4.0
 4.04375
 4.084406316514934
 4.122276374329674
 4.157624929911346
 4.190681774532284
 4.221647746194094
 4.250699481890863
 4.277993217563307
 4.303667858367747
 4.327847482987608
 4.3506434040775765
 4.372155877038263
 ⋮
 4.781262564974172
 4.781872912330582
 4.78245757797536
 4.783017651743386
 4.783554176460985
 4.784068150036249
 4.784560527451245
 4.785032222661161
 4.785484110405099
 4.785917027932993
 4.786331776652848
 4.786729123702271
```



converging to $\sqrt{23} \sim 4.7958$

DOES EVERY STEPSIZE WORK?

E.G., IN THE UNIVARIATE CASE



$$f(x) = \frac{1}{2} x^2$$

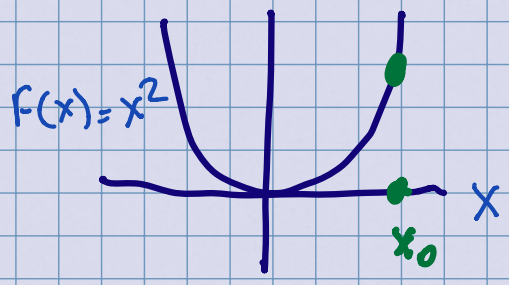
$$\nabla f(x) = x$$

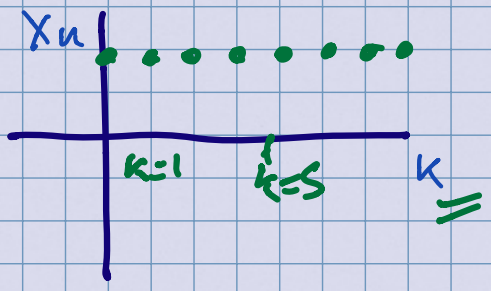
$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

$$x_{k+1} = x_k - \gamma x_k = (1 - \gamma) x_k .$$

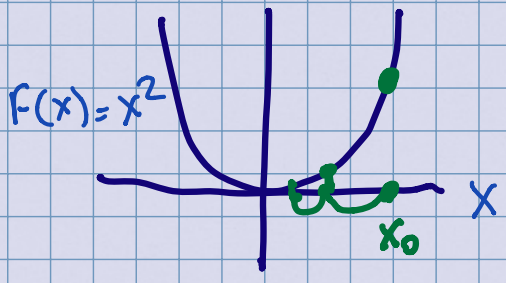
$$|1 - \gamma| < 1$$

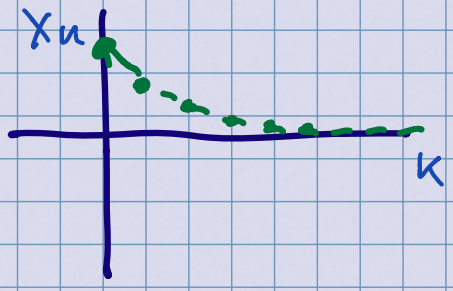
$\gamma = 0$



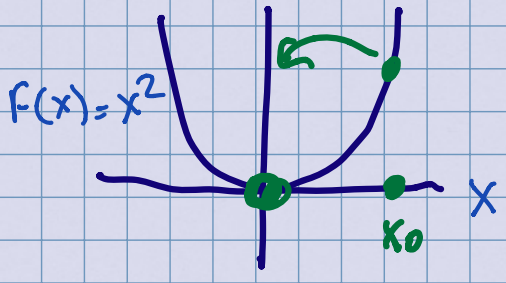


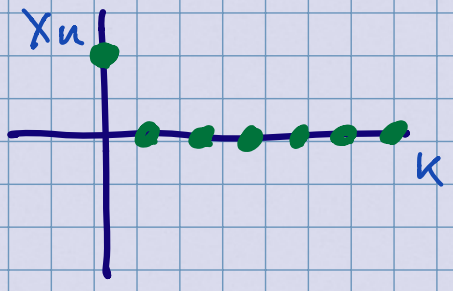
$\gamma = 0.5$



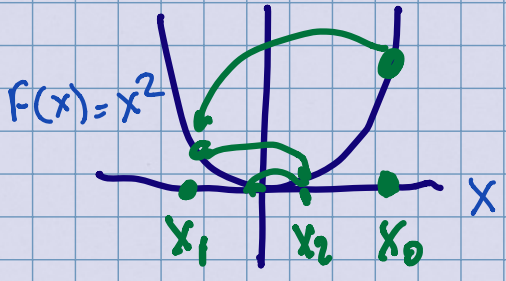


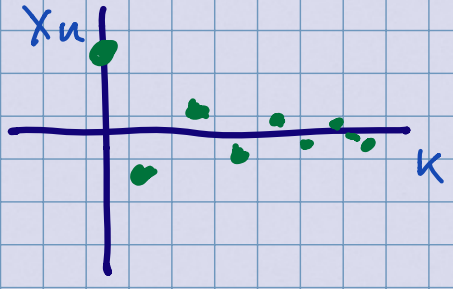
$\gamma = 1$



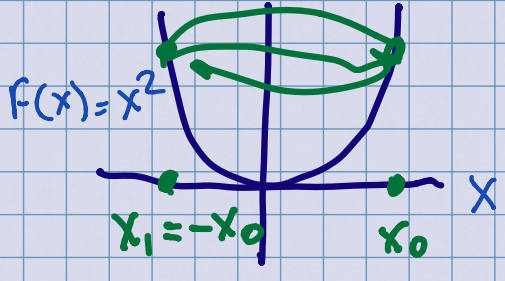


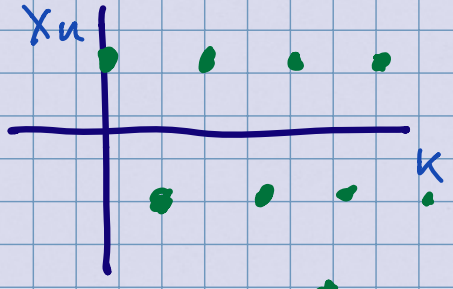
$\gamma = 1.5$



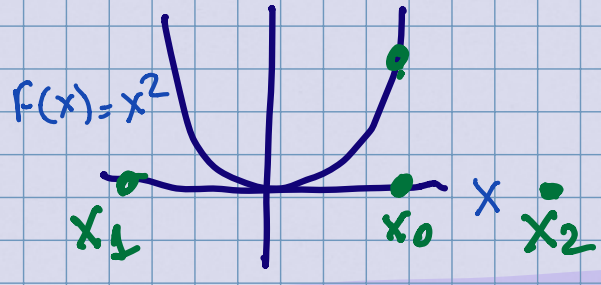


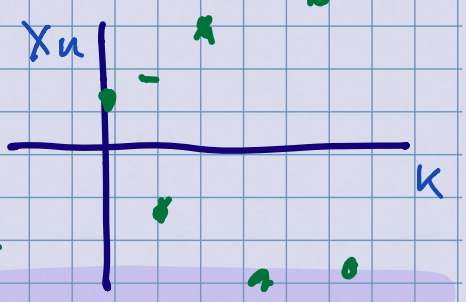
$\gamma = 2$





$\gamma > 2$





What happens for quadratic functions?

$$F(x) = \frac{1}{2} x^T A x$$

(without loss of generality)

A is pd

(so f is strictly convex).

$$\nabla f(x) = Ax$$

eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n > 0$$

$$\Rightarrow x_{k+1} = x_k - \gamma \nabla f(x_k) = x_k - \gamma A x_k$$

$$x_{k+1} = (I - \gamma A) x_k$$

Q: When does this converge?

$$x_{k+1} = M x_k$$

$$M = I - \gamma A$$

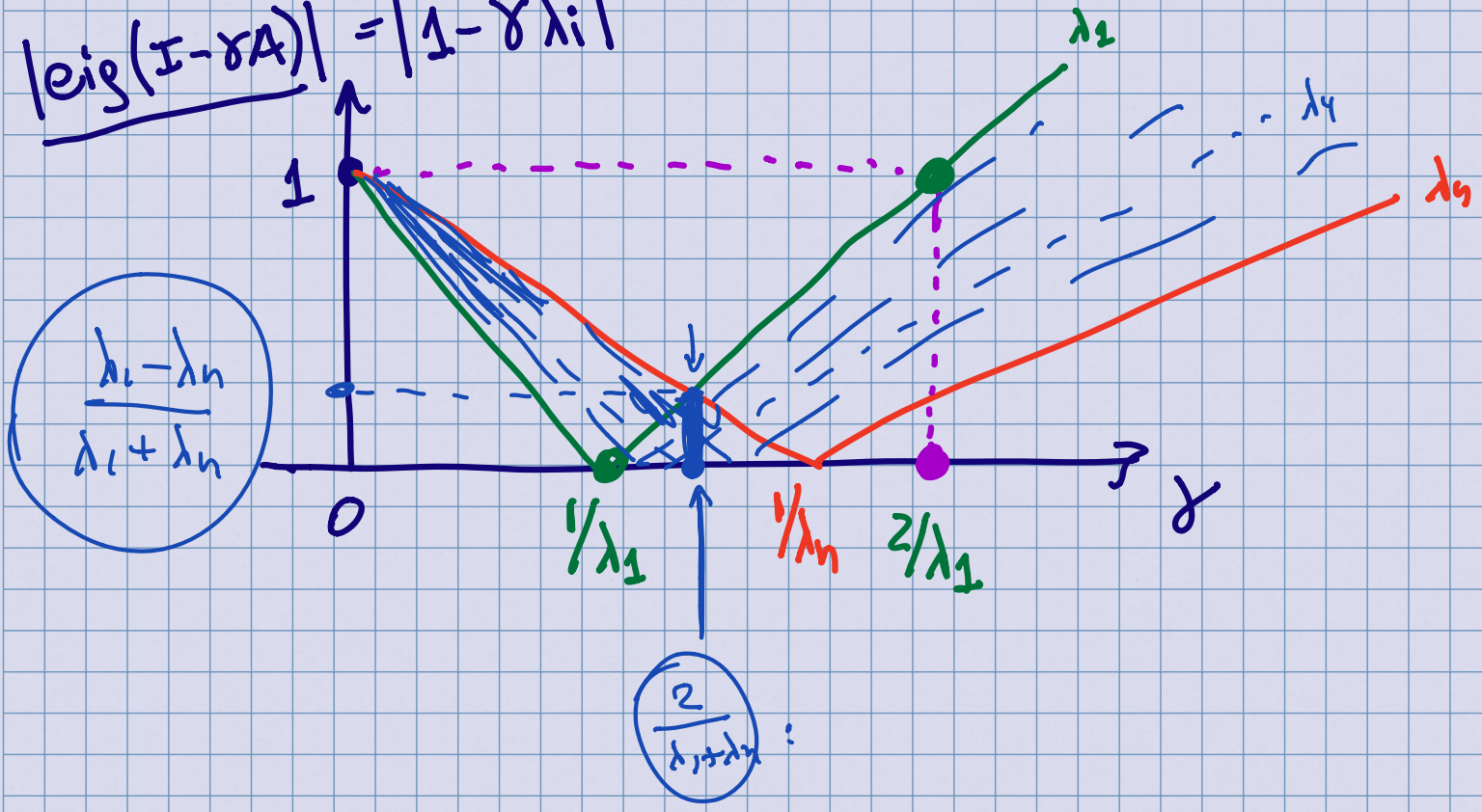
convergence

$$|\lambda_i(M)| < 1$$

Q: What is the "best" stepsize?

$$A = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \qquad I - \gamma A = \begin{pmatrix} 1-\gamma\lambda_1 & & 0 \\ & \ddots & \\ 0 & & 1-\gamma\lambda_n \end{pmatrix}$$

$|\mathrm{eig}(I-\gamma A)| = |1-\gamma\lambda_i|$



$\Rightarrow$ if $\boxed{0 < \gamma < 2/\lambda_1}$

The gradient method converges.

If we measure convergence rate by

$$R = \|I - \gamma A\| = \max_i |1 - \gamma \lambda_i|,$$

"best" stepsize is $\dfrac{2}{\lambda_1 + \lambda_n}$

This choice gives a rate

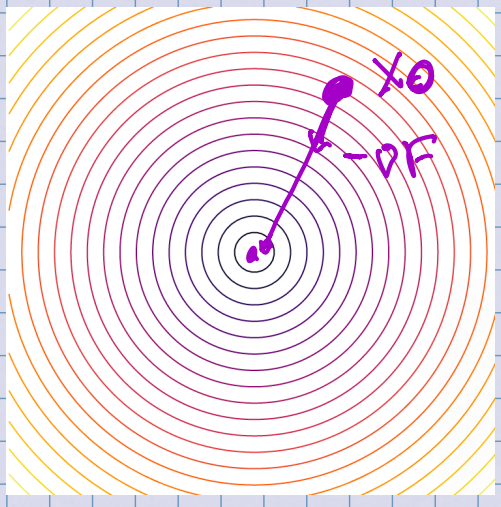
$$R \leq \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = \frac{Q-1}{Q+1}$$

$Q = \dfrac{\lambda_1}{\lambda_n}$ is the condition number of the matrix A.

Behavior in 2D

"zigzag behavior"

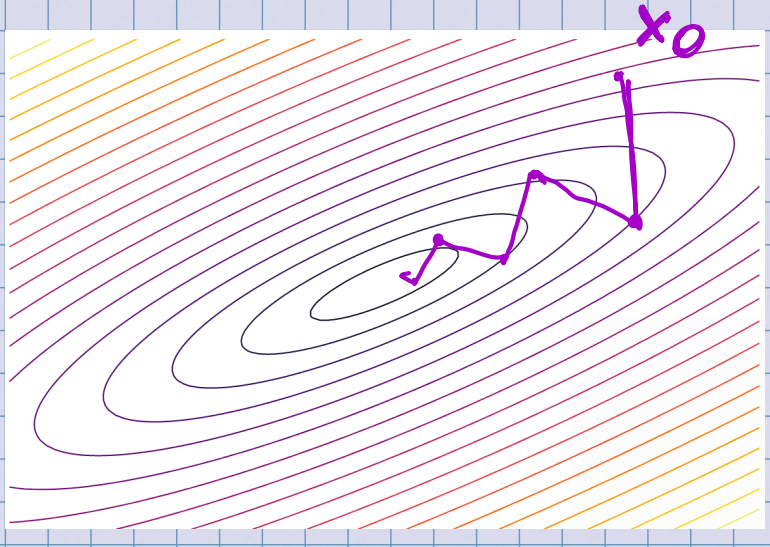
(similar in higher dims)



$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$Q = 1$$

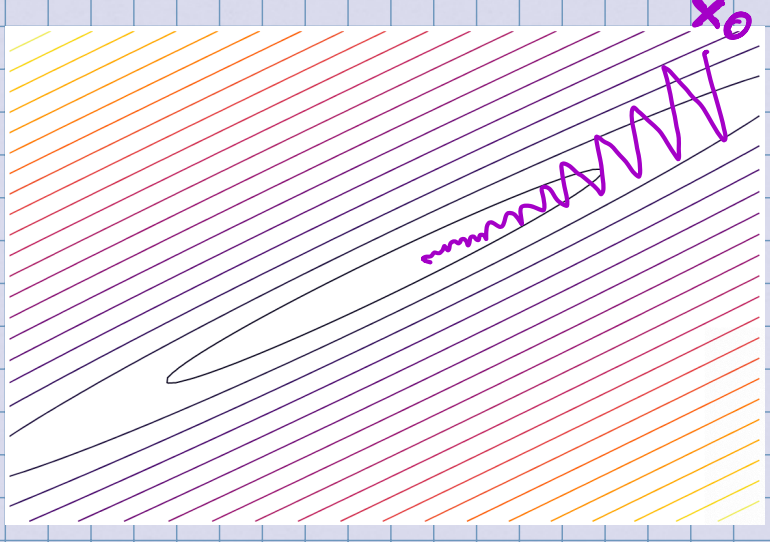
$$f(x,y) = x^2 + y^2$$



$$A = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}$$

$$Q = 10$$

$$F(x,y) = x^2 + 10y^2$$



$$Q = 100$$

What happens in the general convex case?

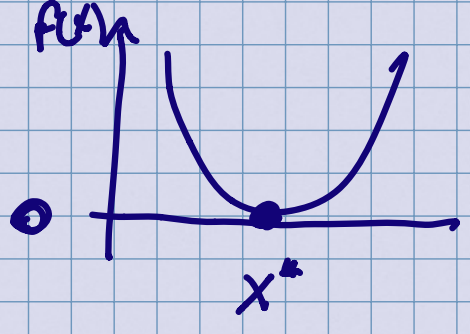
- Does it converge?
- For what values of stepsize?

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

For simplicity

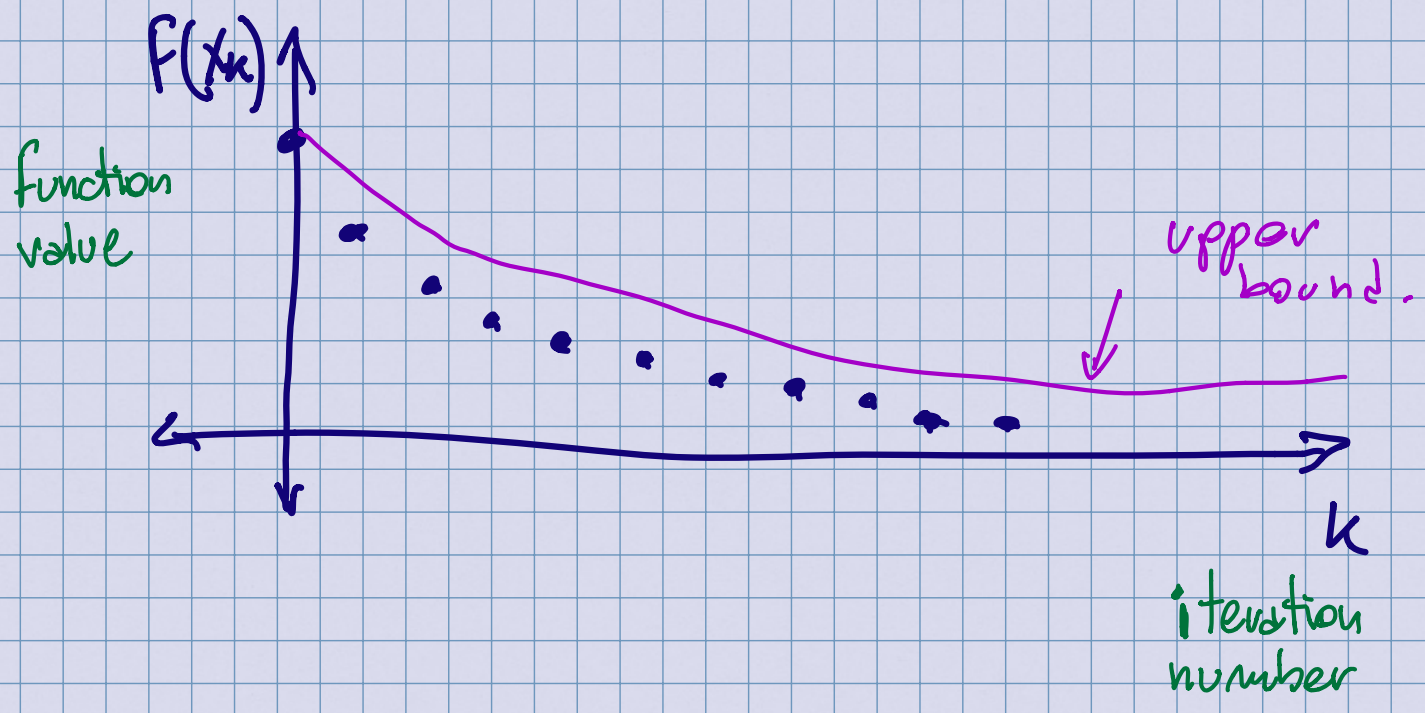
- Assume minimum value is zero:

$$F(x^*) = 0$$



(easy, by shifting the function $F(x) \to F(x) - F(x^*)$)

Q: HOW TO QUANTIFY CONVERGENCE?



Q: WHAT PROPERTIES OF F(x) SHOULD IT DEPEND ON?

Recall:

F(x) CONVEX $\Leftrightarrow$ HESSIAN PSD

(NON NEGATIVE EIGENVALUES)

λ_i eigenvalues of Hessian $\frac{\partial^2 F}{\partial x^2}$

(may depend on the point x)

CASE I: ("Smooth" or "Lipschitz gradient")

λ_i eigenvalues of Hessian

$$0 \leq \lambda_i \leq L$$

(at all points)

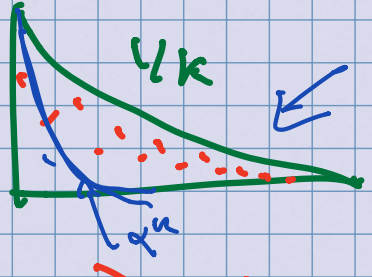
UPPER BOUND ON CURVATURE

Choose Stepsize $\gamma = 1/L$

Then:

$$f(x_k) \leq \frac{L}{2k} \|x_0 - x^*\|^2$$

goes to zero as $k \to \infty$



$\Rightarrow$ GRADIENT DESCENT CONVERGES.

CASE II: ("smooth and strongly convex")

- STRONGER ASSUMPTION: (λ_i are eigenvalues of Hessian)

$$m \leq \lambda_i \leq L \quad \text{(at all points)}$$

UPPER AND LOWER BOUNDS ON CURVATURE

- Choose Stepsize $\gamma = \frac{2}{L+m}$

Then:

$$f(x_k) \leq \frac{L}{2}\left(\frac{Q-1}{Q+1}\right)^{2k} \|x_0 - x^*\|^2$$

where $Q = L/m$ "condition number". $(Q<1)$

$\Rightarrow$ CONVERGES MUCH FASTER!

INTUITION / SUMMARY

- **Stepsize**
  - Too small: slow (but converges)
  - Too large: faster, but may oscillate, or diverge

	FUNCTION PROPERTIES		CONVERGENCE (for best stepsize)	
(I)	SMOOTH	$\lambda_i \leq L$	"slow"	$O(1/k)$
(II)	SMOOTH + STRONGLY CONVEX	$m \leq \lambda_i \leq L$	"fast" or "linear"	$O(\alpha^k)$ (for some $\alpha < 1$)

- no explicit dependence on dimension

- **Condition number** $Q = \frac{\lambda_1}{\lambda_n}$ $\left(\text{or } \frac{L}{m}\right)$

Plays a big role (want Q small)