

# Gradient descent and variants

• Q1: convergence on convex functions

• Def: a set  $K$  is convex

if for all  $x, y \in K$   
 $\lambda x + (1-\lambda)y \in K.$

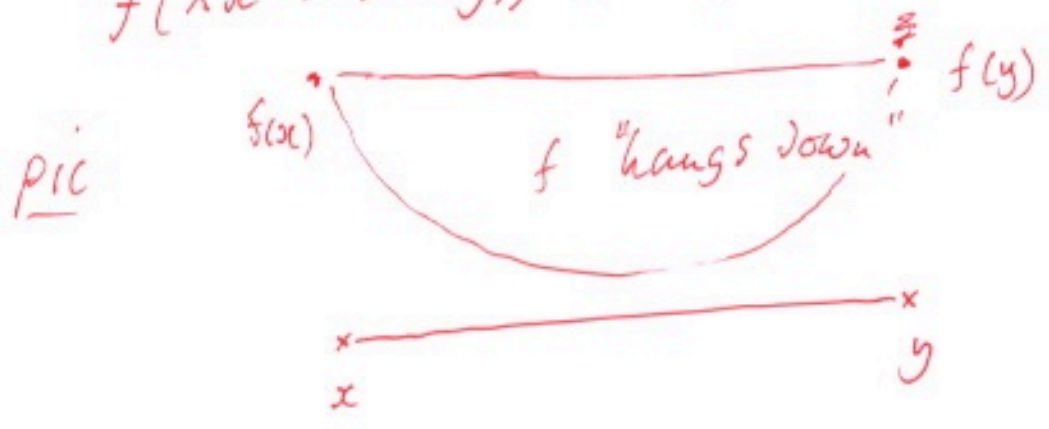
for  $\lambda \in [0, 1]$

(join  $x, y$  by a line segment - its all in  $K$ )

• Def: a function  $f$  is convex

•  $f$  defined on a convex set  $K$ .

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



Another def:

(2)

the epigraph of a function  $f$  is

$$\Sigma(f) = \{(x, y) \mid y \geq f(x)\}$$

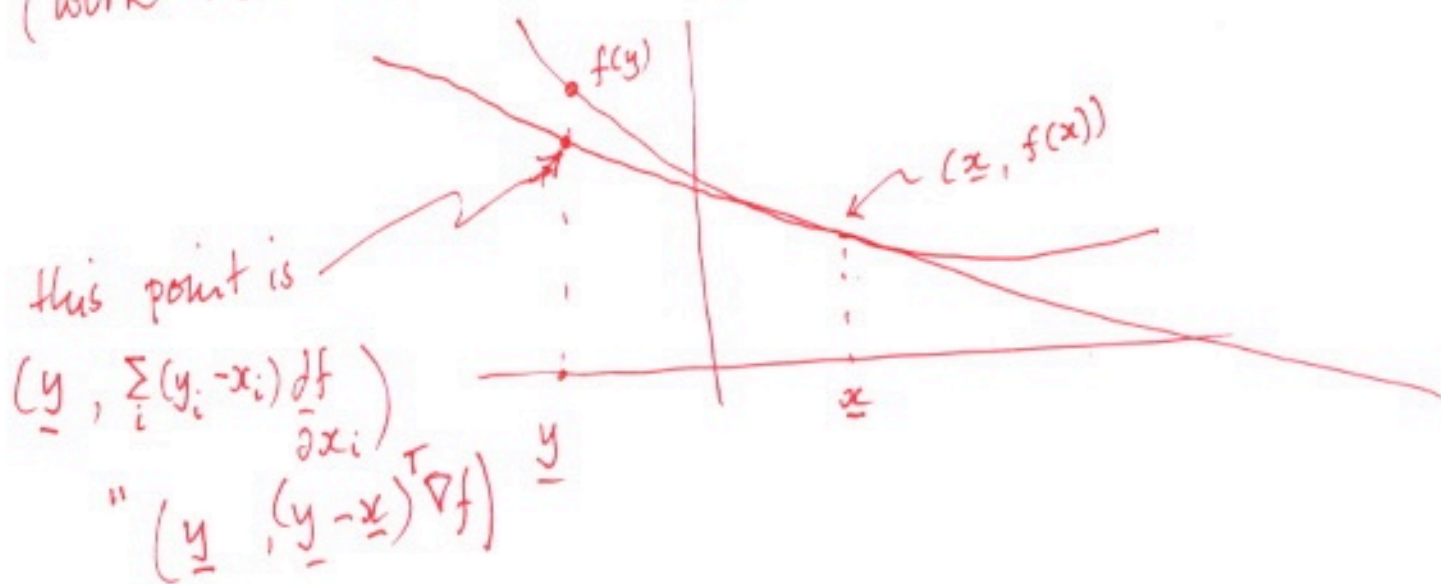


a function is convex  $\Leftrightarrow$  epigraph is convex

if  $f$  is convex AND differentiable

$$f(y) \geq f(x) + \langle \nabla f(x), (y-x) \rangle$$

(work this out with graph)



if  $f$  is  $C^2$  (second derivatives exist everywhere) ③

then

$$f \text{ convex} \iff H_f \succeq 0$$

↑ positive semi-definite

Def:  $K$  convex set;  $f$  a function  
 $f$  is  $G$ -Lipschitz wrt the norm  $\|\cdot\|$

if

$$|f(x) - f(y)| \leq G \|x - y\|$$

Fact: if  $f$  is differentiable

$$f \text{ } G\text{-Lipschitz} \iff \|\nabla f\|_2 \leq G$$

# Gradient Descent.

(4)

$x_0$  starting point

for  $t = 1: T$

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

return  $\hat{x} = \frac{1}{T} \sum_{i=1}^T x_i$

notice this average - helps proof, etc  
but we don't usually do this.

Prop: (G-1) Let  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  be convex, diff, and  $G$ -lipschitz

let  $x^*$  be any point in  $\mathbb{R}^d$

define  $T = \frac{G^2}{\epsilon^2} \|x_0 - x^*\|^2$  and

$\eta = \frac{\|x_0 - x^*\|}{G\sqrt{T}}$  then

Solu  $\hat{x}$  from GD satisfies

$$f(\hat{x}) \leq f(x^*) + \epsilon$$

(in particular, when  $x^*$  is a minimizer)



This comes from

Thm (9.7):

let  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  be convex, diff,  $G$ -lipschitz,

Then gradient descent ensures

$$\sum_{t=1}^T f(x_t) \leq \sum_{t=1}^T f(x^*) + \frac{1}{2} \eta^T G^2 + \frac{1}{2\eta} \|x_0 - x^*\|^2$$

$G=1$  assuming  $G=2$

$$f(\hat{x}) = f\left(\frac{1}{T} \sum_{t=1}^T x_t\right) \leq \frac{1}{T} \sum_{t=1}^T f(x_t)$$

$G=2$  yields

$$\frac{1}{T} \sum_{t=1}^T f(x_t) \leq \underbrace{f(x^*) + \frac{1}{2} \eta G^2 + \frac{1}{2\eta T} \|x_0 - x^*\|^2}_{\text{error}}$$

choose  $\eta = \frac{\|x_0 - x^*\|}{G\sqrt{T}}$

gets :  $f(\hat{x}) \leq f(x^*) + \frac{\|x_0 - x^*\| G}{\sqrt{T}}$

now set

$$T = \frac{1}{\epsilon^2} G^2 \|x_0 - x^*\|^2$$

gets

$$f(\hat{x}) \leq f(x^*) + \epsilon$$

so  $G^{-2} \rightarrow G^{-1}$

Notice :  $x^*$  isn't necessarily close to  $x^*$ ; and we can't show that

imagine function is "flat" close to  $x^*$  we might not get close — notice  $\eta$  gets small as  $T$  ~~goes to~~ gets big.

if we have  $f$  that is strongly convex we can do better.

Proof of G-2

Define  $\phi_t = \frac{\|x_t - x^*\|^2}{2\eta}$

Lemma :

$$f(x_t) + (\phi_{t+1} - \phi_t) \leq f(x^*) + \frac{1}{2}\eta G^2$$

proof :  $\|a+b\|^2 = \|a\|^2 + 2\langle a,b \rangle + \|b\|^2$

(equivalently):  $(a+b)^T(a+b) = a^T a + 2a^T b + b^T b$

so : 
$$\begin{aligned} \phi_{t+1} - \phi_t &= \frac{1}{2\eta} \left[ \|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2 \right] \\ &= \frac{1}{2\eta} \left[ 2(x_{t+1} - x_t)^T(x_t - x^*) + (x_{t+1} - x_t)^T(x_{t+1} - x_t) \right] \end{aligned}$$

but :  $x_{t+1} = x_t - \eta \nabla f(x_t)$

§

So:

$$\phi_{t+1} - \phi_t = \frac{1}{2\eta} \left[ 2(-\eta \nabla f(x_t))^T (x_t - x^*) + \|\eta \nabla f(x_t)\|^2 \right] \quad \textcircled{8}$$

but  $f$  is  $G$ -Lipschitz, so

$$\phi_{t+1} - \phi_t = [\nabla f(x_t)]^T [x_t - x^*] + \frac{1}{2} \eta G^2$$

$$f(x_t) + \phi_{t+1} - \phi_t \leq f(x_t) + [\nabla f(x_t)]^T [x_t - x^*] + \frac{1}{2} \eta G^2$$

but

$$f(x_t) + [\nabla f(x_t)]^T [x_t - x^*] \leq f(x^*)$$

So

$$f(x_t) + \phi_{t+1} - \phi_t \leq f(x^*) + \frac{1}{2} \eta G^2$$



Proof of G-2:

9

Start with

$$f(x_t) + \phi_{t+1} - \phi_t \leq f(x^*) + \frac{1}{2} \eta G^2$$

Sum over  $t$

$$\sum_1^T f(x_t) + \sum_1^T (\phi_{t+1} - \phi_t) \leq T f(x^*) + \frac{1}{2} \eta G^2 T$$



$$\phi_T - \phi_1$$

but  $\phi_T \geq 0$

So

$$\frac{1}{T} \sum_1^T f(x_t) - \frac{1}{T} \phi_1 \leq f(x^*) + \frac{1}{2} \eta G^2$$

So

$$\frac{1}{T} \sum_1^T f(x_t) \leq f(x^*) + \frac{1}{2} \eta G^2 + \frac{\|x_1 - x^*\|^2}{2\eta T}$$

QED

# Stronger assumptions yield better behavior (10)

Def: strong convexity

$f: K \rightarrow \mathbb{R}$  is  $\alpha$ -strongly convex if any of these holds

1) (No deriv)

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \left(\frac{\alpha}{2}\right)\lambda(1-\lambda)\|x-y\|^2$$

all  $\lambda \in [0, 1]$

2) (diff)

$$f(y) \geq f(x) + [\nabla f]^T [y-x] + \frac{\alpha}{2} \|x-y\|^2$$

3) (2 diff)

‡ smallest eigenvalue of  $H_f \geq \alpha$   
at all points

↑ all of these mean: the function "grows" faster than some  $\alpha$ .

Def: Lipschitz smoothness

(11)

$f: K \rightarrow \mathbb{R}$  is  $\beta$ -Lipschitz-smooth if any of these hold

1) (no deriv)

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y) - \frac{\beta}{2} \lambda(1-\lambda) \|x-y\|^2$$

all  $\lambda \in [0,1]$

2) (1st)

$$f(y) \leq f(x) + [\nabla f]^T [y-x] + \frac{\beta}{2} \|x-y\|^2$$

3) (2nd)

all eigenvalues of  $H_f \leq \beta$  at every point

All of these mean: function doesn't  
"grow too fast"

taken together, these allow better bounds (expected - the derivative at a point tells us more about function at some other point) (12)

Thm:  $f$   $\beta$ -smooth AND  $\alpha$ -strongly convex

$x^*$  soln to  $\arg\min_{x \in \mathbb{R}^n} f(x)$ .

then running gradient descent w/  $\eta = \frac{1}{\beta}$

gives

$$f(x_t) - f(x^*) \leq \frac{\beta}{2} e^{\left(-\frac{t}{k}\right)} \cdot \|x_1 - x^*\|^2$$

$k = \text{cond number}$   
 $= \beta/\alpha$

(Proof in notes)



Because the function is  $\alpha$ -strongly convex, (13)

$$f(x_t) \geq f(x^*) + [\nabla f]^T [y - x] + \frac{\alpha}{2} \|x_t - x^*\|^2$$

$x^*$  is min

so:  $\left[ f(x_t) - f(x^*) \right] \frac{2}{\alpha} \geq \|x_t - x^*\|^2$

↑

we can bound how close we are to right answer.

$$f(x_t) - f(x^*) \leq \text{Const } e^{-t/k}$$

and bound above means

$$\|x_t - x^*\|^2 \leq \text{Const. } e^{-t/k}$$

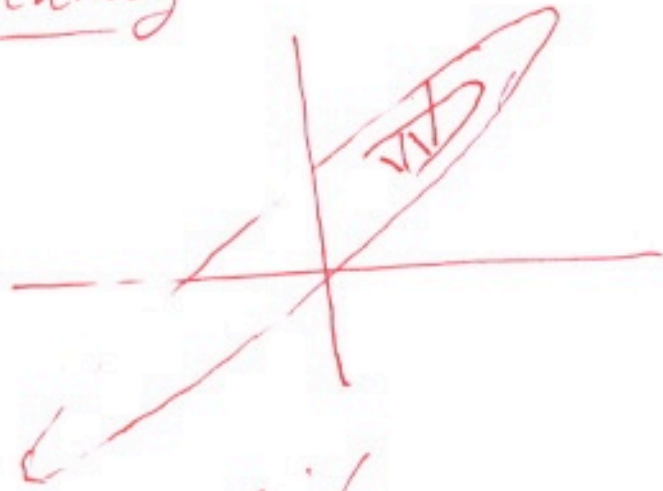
which is often interpreted as:

the number of  $\left. \begin{array}{l} \text{bits of precision} \\ \text{digits of solu} \end{array} \right\}$

known is linear in number of iterations

Discontents:

- Poor scaling



(we've seen this!)

Symptoms:

- components of gradient switch sign a lot
- poor progress.

- Expensive iterations:

$$f(x) = \frac{1}{N} \sum_{i=1}^N \ell_i(x) + g(x)$$

with big  $N$

## Poor scaling strategies:

- Newton or QN ideal but can't get Hessian
- adjust gradient.

## Example:

$$\frac{1}{2} [x^2 + \epsilon y^2] \quad 0 < \epsilon \ll \frac{1}{2}$$

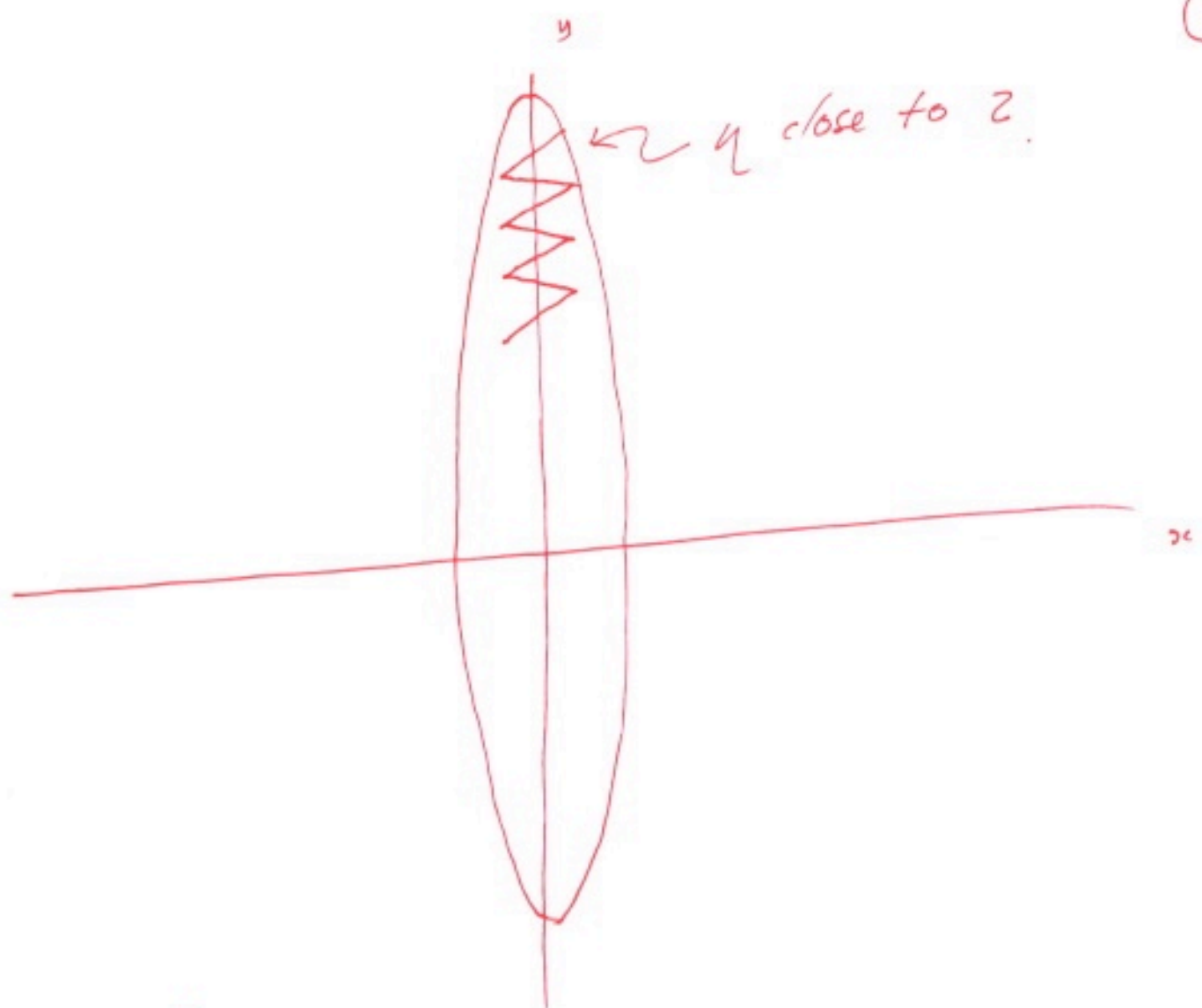
$$\begin{matrix} x_{i+1} &= & x_i & - \eta & x_i \\ y_{i+1} &= & y_i & - \eta & \epsilon y_i \end{matrix} \quad \text{so } \begin{matrix} x_i &= & (1-\eta)^i x_0 \\ y_i &= & (1-\epsilon\eta)^i y_0 \end{matrix}$$

so convergence when  $0 < \eta < 2$

cases

$\eta < 1$	$\approx$ slow!
$\eta = 1$	$x$ 1-step, $y$ slow travel
$\eta = 2-\delta$	$\leftarrow$ slow, but faster; $x$ changes sign at <u>every step</u>





Momentum

$$m_i = (1-\alpha) m_{i-1} + \alpha \nabla f(x_{i-1})$$

$$x_i = x_{i-1} - \eta m_i$$

IDEA :  $\cdot m_i$  is a moving average of gradients  
 $\cdot$  this should smooth.

# Strategies for expensive eval

18

$$f(x) = \frac{1}{N} \sum_i h_i(x) + g(x)$$

Notice  $\frac{1}{N} \sum_i h_i(x)$  is a population mean

~~we~~ IDEA draw a sample of  $s$  items and form the sample mean.

$$\frac{1}{s} \sum_{u=1}^s h_u(x) = m_s(x) \leftarrow \begin{array}{l} \text{mean over sample} \\ \text{this is a random var} \end{array}$$

Assume ~~some~~  $h_u$  are drawn uniformly at random with replacement

Then

$$\begin{aligned} E[m_s(x)] &= \frac{1}{N} \sum_i h_i(x) \\ \text{Var}[m_s(x)] &= \frac{1}{N} \text{const} \end{aligned} \leftarrow \begin{array}{l} \text{usually hard} \\ \text{to compute} \end{array}$$

IDEA:

(19)

• compute  $\nabla_x m_s(x)$   
and use this as gradient

• Simple analysis

• We're going in the right direction on average, so ...

• More sophisticated analysis requires new ideas