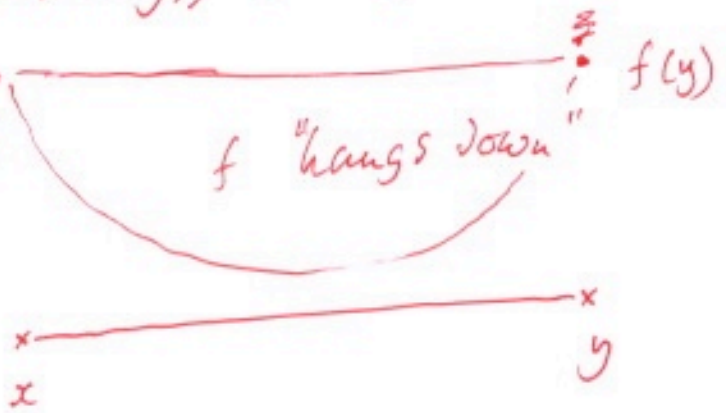
Gradient descent and variants

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

Pic



Another def:

the epigraph of a function f is

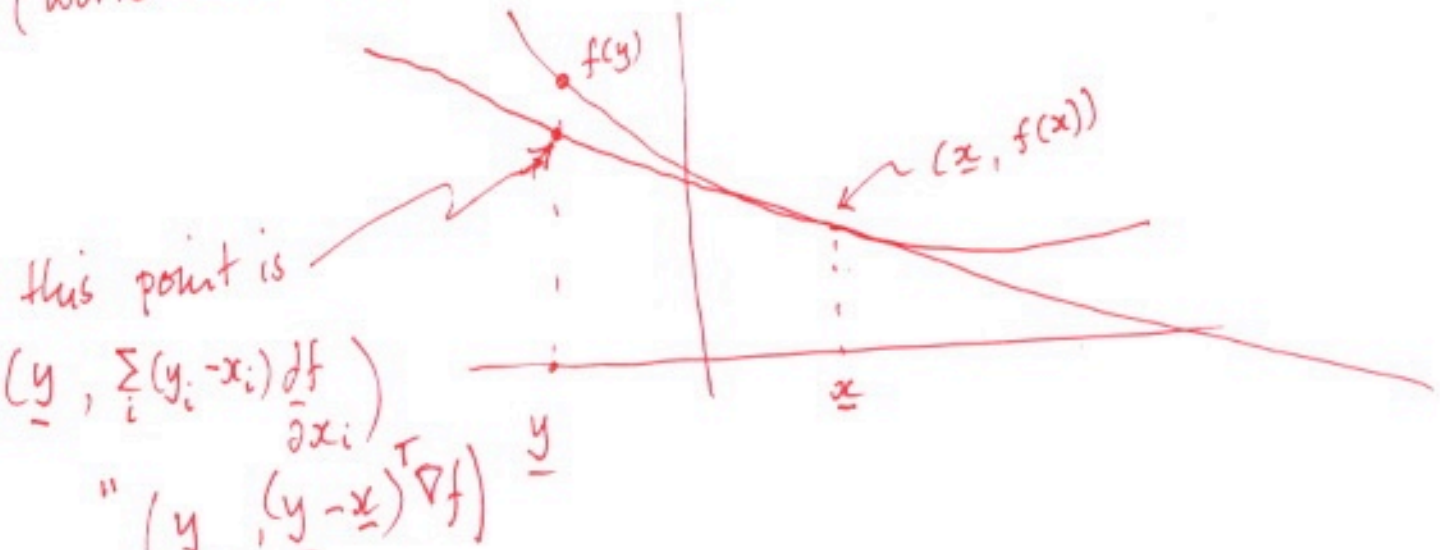
$$\mathcal{E}(f) = \{(x, y) \mid y \geq f(x)\}$$

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)

then

f convex $\iff$ $H_f \succeq 0$

↑ positive semi definite

Def: K convex set, f a function

f is G-Lipschitz wrt the norm $\|\ \|$

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

Fact: if f is differentiable

f G-Lipschitz $\iff \|\nabla f\|_2 \leq G$

Gradient Descent.

x_1 starting point

for $t = 1 : T$

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

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

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

Prop: (G-1) Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, diff, and G-lipschitz. Let x^* be any point in $\mathbb{R}^d$

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

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

Soln $\hat{x}$ from GD satisfies

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

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

This comes from

Thm (4-2):

Let $f: R^d \to R$ 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

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

G-2 yields

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

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

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

now set

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

gets

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

so $G-2 \rightarrow G-1$

Notice: $\hat{x}$ 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
η gets small as T ~~goes etc~~ 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:

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

(equiv: $(a+b)^T(a+b) = a^Ta + 2a^Tb + b^Tb$)

So:

$$\phi_{t+1} - \phi_t = \frac{1}{2\eta}\left[\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right]$$

$$= \frac{1}{2\eta}\left[2(x_{t+1} - x_t)^T(x_t - x^*) + (x_{t+1} - x_t)^T(x_{t+1} - x_t)\right]$$

but:

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

So:

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

but f is G-Lipschitz, so

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

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

but

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

So

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

Proof of G-2:

Start with

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

Sum over t

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

$$\updownarrow$$

$$\phi_T - \phi_1$$

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

Def: strong convexity

$f: K \to R$ is α-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)

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

3) (2 Diff)

$\nabla^2 f$ = smallest eigenvalue of $H_f \geq \alpha$ at all points

↑ all of these mean: the function "grows" faster than some α.

Def: Lipschitz smoothness

$f: K \to R$ is β-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) (1 diff)

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

3) (2 diff)

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)

Thm: f β-smooth AND α-strongly convex

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

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

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

κ = cond number = β/α

(Proof in notes)

Because the function is α-strongly convex,

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

x^* is min $\Rightarrow$ 0

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

↑ 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

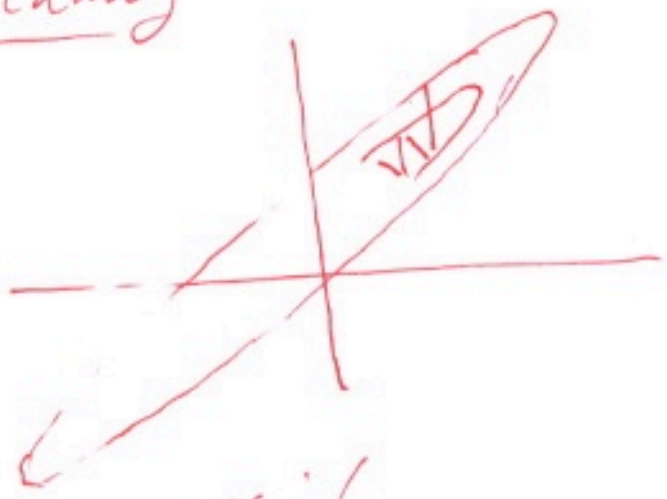
$$\|x_t - x^*\|^2 \leq \text{Const. } e^{-t/k}$$

which is often interpreted as:

the number of { bits of precision / digits of soln } 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:

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

with big N

Poor scaling strategies:

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

Example:

$$\frac{1}{2}\left[x^2 + \varepsilon y^2\right] \qquad 0 < \varepsilon \ll \frac{1}{2}$$

$$\begin{matrix} x_{i+1} \\ y_{i+1} \end{matrix} = \begin{matrix} x_i \\ y_i \end{matrix} - \eta \begin{matrix} x_i \\ \varepsilon y_i \end{matrix}$$

so $x_i = (1-\eta)^i x_0$

$y_i = (1-\varepsilon\eta)^i y_0$

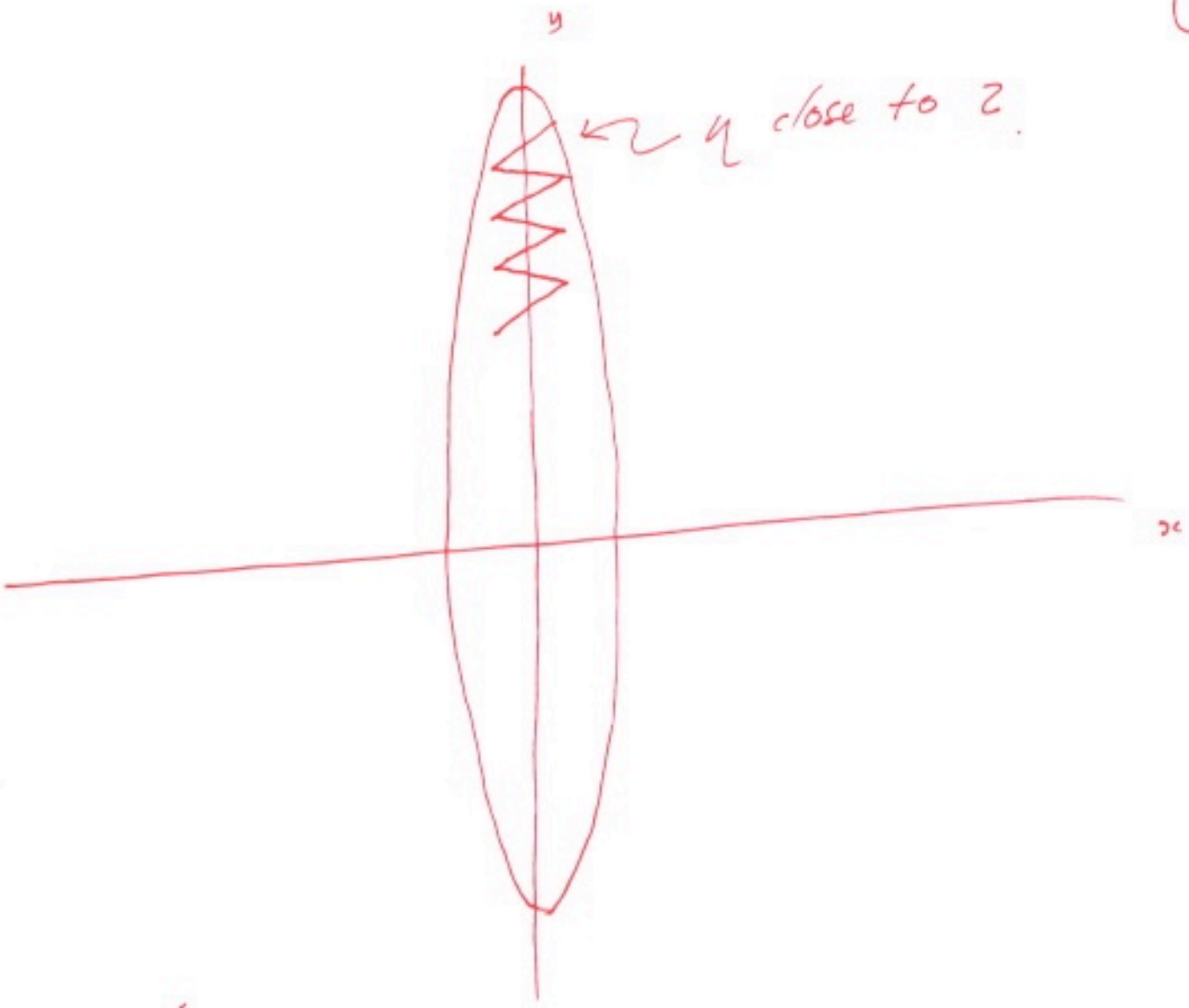
so convergence when $0 < \eta < 2$

cases $\eta < 1$ ← slow!

x 1-step, y slow travel

$\eta = 1$

$\eta = 2-\delta$ ← slow, but faster;
x changes sign at every step



momentum

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

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

IDEA: . m_i is a moving average of gradients
. this **should** smooth.

Strategies for expensive eval

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

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

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

$$\frac{1}{S}\sum_{u=1}^{S} L_u(x) = M_S(x)$$ ← mean over sample
this is a random var

Assume L_u are drawn uniformly at random with replacement

Then

$$E[M_S(x)] = \frac{1}{N}\sum_i L_i(x)$$ ← usually hard to compute

$$\mathrm{Var}[M_S(x)] = \frac{1}{N}\,\mathrm{Const}$$

IDEA:

- compute $\nabla_x u_s(x)$ and use this as gradient
- Simple analysis
  - We're going in the right direction on average, so ...
- More sophisticated analysis requires new ideas

step length (= learning rate) presents issues

- Simple example

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

$$\nabla_x f = \left[\frac{1}{N} \sum_i a_i\right] x = \bar{a}\, x$$

- GD: (without averaging!)

x_0 - start

$$x_{\omega+1} = x_\omega - \eta \nabla f = (1 - \eta \bar{a}) x_\omega$$

$$= (1 - \eta \bar{a})^{\omega-1} x_0$$

- convergence requires $0 < \eta \bar{a} < 2$
- ~~and $\eta \bar{a} \neq 1$~~
- and we would like $|1 - \eta \bar{a}|$ close to zero, for speed
- hard if we don't know $\bar{a}$

$$g_\omega = \bar{a} x_\omega + \xi_\omega$$

$\bar{a} x_\omega$: expected value of gradient at x_ω

ξ_ω: random var, mean = 0, var = $\frac{c}{M}$ (M = number of samples)

$$x_{\omega+1} = x_\omega - \eta g_\omega = x_\omega - \eta \bar{a} x_\omega - \eta \xi_\omega$$

$$= (1-\eta\bar{a})^{\omega+1} x_0 - \eta \left[(1-\eta\bar{a})^{\omega} \xi_0 + (1-\eta\bar{a})^{\omega-1} \xi_1 + \cdots + \xi_\omega \right]$$

$$= (1-\eta\bar{a})^{\omega+1} x_0 - \eta \left[\frac{1 + |(1-\eta\bar{a})|^{\omega+1}}{-\eta\bar{a}} \right] \xi_0$$

this works because (a) geometric series and (b) ξ_ω and $\xi_{\omega-1}$ etc are independent, so variances add (works only if M is small compared to N)

we are dealing with a random error · with · var

$$\left| \left[\frac{1 + |(1 - \eta \cdot \bar{a})|^{w+1}}{\bar{a}} \right] \right|$$

(so if η too big, we're in trouble)
η too small, same

Strategy (we don't know $\bar{a}$!)

- est $\bar{a}$ [? how ?]
- choose some fixed η
  - take a lot of steps
  - now reduce η
    - take a lot of steps
    - etc

When? by how much?

Seat of pants