Some convergence results

consider f, convex, with first derivative and 1st derivative is Lipschitz with const L:

i.e. $\|f'(x) - f'(y)\| \leq L\|x-y\|$

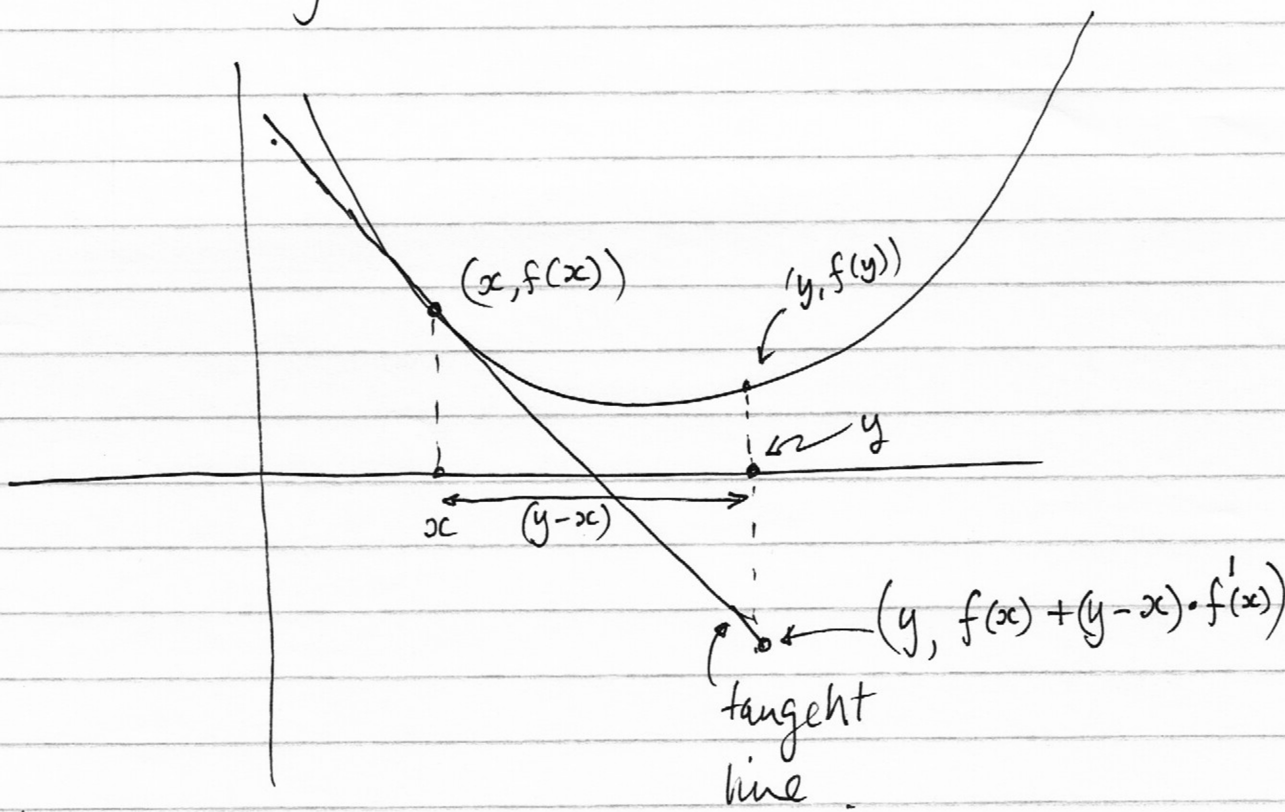
Some useful inequalities follow:

Ⓐ $0 \leq f(y) - f(x) - f'(x)\cdot(y-x) \leq \frac{L}{2}\|y-x\|^2$

Ⓑ $\frac{1}{L}\|f'(x) - f'(y)\|^2 \leq (f'(x) - f'(y))\cdot(x-y)$

Proof of A (instructive)

f convex, so that graph of f always lies above its tangent line



equivalently

$$f(y) \geqslant f(x) + (y-x)\cdot f'(x)$$

or $$0 \leqslant f(y) - f(x) - (y-x)\cdot f'(x).$$

(first side of A)

Second step of A is also instructive

Interpret

$f(x) + (y-x) \cdot f'(x)$

as: value of f predicted at y, assuming derivative of f is constant at the value $f'(x)$

so second $\leq$ bounds the difference between this and true value at y.

→ this should work because

$\| f'(y) - f'(x) \| \leq L \| x - y \|$

"Sloppy" proof:

- Assume that the derivative changes as fast as possible from x to y

- so $f'(u) = f'(x) + L(u-x)$

(signs don't matter)

- then

$$f(y) = f(x) + \int_x^y f'(u)\,du$$

$$= f(x) + (y-x)f'(x) + \int_x^y L(u-x)\,du$$

$$= f(x) + (y-x)f'(x) + \frac{L}{2}\left[y^2-x^2\right] - Lx\left[y-x\right]$$

$$= f(x) + (y-x)f'(x) + \frac{L}{2}\left[y-x\right]^2$$

now restore ineq, and we are done.

From this, we can get.

alg:
$$x_{k+1} = x_k + h\left(-f'(x_k)\right).$$

then

$$f(x_k) - f^* \leq O\left(\frac{1}{k}\right)$$

(f^* = value at optimal point)

Proof (ish!)

write

location of opt.

$$r_k = \| x_k - x^* \|$$

$$r_{k+1}^2 = \| x_k - x^* - h f_k' \|^2$$
$$= r_k^2 - 2h f_k' \cdot (x_k - x^*) + h^2 \| f'(x_k) \|^2$$

now $f'(x^*) = 0$, so

$$f'_k \cdot (x_k - x^*) = (f'_k - f'(x^*)) \cdot (x_k - x^*)$$
$$\geq \frac{1}{L} \|f'_k - f'(x^*)\|^2$$
$$= \frac{1}{L} \|f'_k\|^2$$

So $$r_{k+1}^2 \leq r_k^2 - h\left(\frac{2}{L} - h\right) \|f'_k\|^2$$

[Notice this works because $\|f'_k\|^2$ can't be large for large enough k.]

so $$r_k \leq r_0 .$$

Now (A) gives →

$$f(x_{k+1}) \leqslant f(x_k) + f'_k \cdot (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

$$= f(x_k) + f'_k \cdot (h[-f'_k]) + \frac{L}{2} h^2 \|f'_k\|^2$$

$$= f(x_k) - \omega \|f'_k\|^2$$

$h(1 - \frac{L}{2}h)$

Now write $\Delta_k = f(x_k) - f^*$

~~then~~

Now ~~Δ_k~~

$$f^* \geqslant f(x_k) + f'_k \cdot (x^* - x_k)$$

so $$f'_k \cdot (x_k - x^*) \geqslant f(x_k) - f^*$$

so $$\Delta_k \leqslant f'_k \cdot (x_k - x^*)$$

$$= f'_k \cdot r_k$$

$$\leqslant \|f'_k\| \cdot r_0$$

recall $r_0 \geqslant r_k$

so

$$f(x_{k+1}) - f^* \leqslant (f(x_k) - f^*) - \omega \|f'_k\|^2$$

we have

$$\left(\frac{\Delta_k}{r_0}\right)^2 \leqslant \|f_k\|^2$$

Notice $-$ sign, and get.

$$f(x_{k+1}) - f^* = \Delta_{k+1}$$

$$\leqslant \Delta_k - \frac{\omega}{r_0^2} \Delta_k^2$$

so

$$\Delta_{k+1} \leqslant \Delta_k \left[1 - \frac{\omega}{r_0^2} \Delta_k \right]$$

so:

$$\frac{\Delta_{k+1}}{\Delta_k} + \frac{\omega}{r_0^2} \Delta_k \leqslant 1$$

So:

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{\Gamma_0^2} \cdot \frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{\Gamma_0^2}$$

So:

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_0} + \frac{\omega}{\Gamma_0^2} \cdot (k+1).$$

So

$$\Delta_{k+1} \leq O\left(\frac{1}{k}\right)$$

$\longrightarrow$ QED!

Strongly convex fns

Recall a convex fn has

$$f(y) \geq f(x) + f'(x) \cdot (y-x)$$

f is strongly convex if

$$f(y) \geq f(x) + f'(x) \cdot (y-x) + \frac{\mu}{2} \|x-y\|^2$$

- this term guarantee's growth
- $\mu > 0$; value of μ matters

Particularly interesting are:

- Strongly constant fn's w/ Lipschitz first derivative.

Thm: for $f \in S^{1,1}_{\mu,L}$

$\Big\{$ strongly convex w/ μ;
1-diff
1st der. Lipschitz, L.

then:

$$\left(f'(x) - f'(y)\right) \cdot (x-y) \geq \frac{\mu L}{\mu + L} \|x-y\|^2 + \frac{1}{\mu+L} \|f'(x) - f'(y)\|^2 .$$

Proof:

write $\phi(x) = f - \frac{\mu}{2}\|x\|^2$

then $\phi' = f' - \mu x$

so: ϕ is convex, ϕ has Lipschitz first d, const $L-\mu$.

~~Gx~~ Case: $\mu < L$

by B, we have

$$\frac{1}{(L-\mu)} \|\phi'(x) - \phi'(y)\|^2 \leq (\phi'(x) - \phi'(y)) \cdot (x-y)$$

rearrange to get original $\square$

Thm:

if $f \in S^{1,1}_{\mu,L}$, $0 < h < \frac{2}{\mu+L}$

then gradient method generates x_k

st:

$$\|x_k - x^*\|^2 \leq \rho^k \|x_0 - x^*\|^2$$

true min

Proof

write

$$r_K = \| x_K - x^* \|$$

$$r_{K+1}^2 = \| x_K - x^* - h f_K' \|^2$$

$$= r_K^2 - 2h f_K' \cdot (x_K - x^*) + h^2 \| f_K' \|^2$$

Notice $f'(x^*) = 0$

so $$(f_K' - f_*') \cdot (x_K - x^*) \geqslant \frac{\mu L}{\mu + L} \| x_K - x^* \|^2 + \frac{1}{\mu + L} \cdot \| f_K' - f_*' \|^2$$

(by above)

So:

$$r_{k+1}^2 \leq r_k^2 - 2h\left[\frac{\mu L}{\mu+L} \cdot r_k^2 + \frac{1}{\mu+L} \|f_k'\|^2\right]$$

$$+ \quad h^2 \|f_k'\|^2$$

$$= \left(1 - \frac{2h\mu L}{\mu+L}\right) r_k^2 + h\left(h - \frac{2}{\mu+L}\right) \|f_k'\|^2$$

Now there is a step in Nesterov 2004, 2.1.15, that I don't follow

Stochastic gradient descent:

- write objective

$$f = \frac{1}{N}\sum_i f_i$$

- choose one f_i, UAR. say $\cancel{} \, e$

- $$x_{k+1} = x_k - \alpha_k \nabla f_e.$$

Notice:

∇f_e is a randomized estimate of ∇f

$$E[\nabla f_e] = \nabla f$$

In practice, this is rather well behaved with reasonable step-length schedules

- Always assume subgradient bounded, opt. soln. exists.

Thm: (Nedic + Bertsekas)

assume $\lim\limits_{k \to \infty} \alpha_k = 0$ and $\sum\limits_{k=0}^{\infty} \alpha_k = \infty$, f convex

Then

$$\lim_{k \to \infty} \left[\inf f(x_k) \right] = f^*$$

(i.e. eventually gets to the right place)

Thm (N+B)

Assume $$\sum_k \alpha_k = \infty$$

$$\sum_k \alpha_k^2 < \infty$$

Then x_k converges to an optimal soln

Thm for f convex

$$E[f\cancel{g}(x_k)] - f(x^*) = O(1/\sqrt{k})$$

(Nemirovski, '09)

Thm: for f strongly convex

$$E[f(x_k)] - f(x^*) = O(1/k)$$

(Nemirovski 09)

In Practice:

~~Step~~

- Quite well-behaved in early stages (large, useful steps)
- Can be slow in late stages
- steplength schedules are a major nuisance
- Convergence diagnosis is hopeless

Q: Why is convergence slow?

A: The estimate of ∇f is noisy

Strategies

- filter, smooth, average, etc.

eg weighted average of all past gradients

- write g_k for grad. est at k.

$$g_{k+1} = (1-\alpha)\nabla f_e + \alpha g_k$$

(Notice this never forgets old gradients) eg. use more samples

g_K = [Ave over r samples]

but notice the SD of this est goes down as $1/\sqrt{r}$,

so diminishing returns.

eg. Momentum

$$x_{K+1} = x_K - \alpha_K f_l'(x_K) + \beta_K (x_K - x_{K-1})$$

Usual to use $\beta_K = \beta$.

In this case,

$$x_{k+1} = x_k - \underbrace{\sum_{j=1}^{k} \alpha_j \cdot \beta^{k-j} \cdot \nabla f_{\ell_j}(x_j)}$$

- geometrically weighted average of all previous steps.

- problem
  - "large" β $\equiv$ no forgetting
  - "small" β $\equiv$ no point.

SAG

- Pretend we can maintain an array of gradients, 1 per term in the sum.

$$y_{ik}$$

$\uparrow$ term $\quad \uparrow$ iteration

- Initialize w/ $y_{ik} = 0$
- Step:
  - select l UAR $\in [1 \cdot N]$
  - $y_{ik} = \begin{cases} \nabla f_l(x_k) & \text{for } i = l \\ y_{ik-1} & \text{otherwise} \end{cases}$

$$\cdot \quad x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_i y_{ik}.$$

Notice that

$$\frac{1}{n} \sum_i y_{ik} \quad \text{is an estimate}$$

of the gradient, where one always uses most recent grad. $\Rightarrow$ no issue of "forgetting"

Convergence:

- for f_i convex, diff., f' Lipschitz, constant L.
- $\bar{x}_K = \frac{1}{K} \sum_{i=1}^{K} x_i$

(I) $E[f(\bar{x}_K)] - f(x^*) = O(\frac{1}{K})$

for appropriate constant steplength.
(cf. grad. descent).

(II) f μ-strongly convex gives

$$E[f(x_K)] - f(x^*) \leq \rho^K C_0$$

(again constant steplength).

Notice that 2 can be applied to iterates, because

$$\frac{\mu}{2} \| x_k - x^* \| \le f(x_k) - f(x^*)$$

from strong convexity.

<u>Simple pseudo code</u>

$d = 0$; $y_i = 0$ for $i = 1 \cdots n$

for $k = 0 \cdots$

- Sample i UAR from $[1, n]$
- $d = d - y_i$
- $y_i = \nabla f_i$
- $d = d + y_i$
- $x = x - \frac{\alpha}{n} d$

end.

Issues

best steplength is $\frac{1}{16L}$.

- but what if we don't know L?
  - start with an initial estimate L_0
  - at each step, can check.

    because we know f' at

    two points
  - if L_0 too small

    $L_0 \to 2 * L_0$

- Convergence diagnosis is straightforward
  - look at $\|d\|$

- Experience, convergence results suggest
  - if we can afford only one pass through data, SG

SAG.

  - if we can afford 100's of passes, FG