GD1 SUMMARY of GRADIENT DESCENT COMPLEXITY

Note the difference!

Strict Convexity $f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$

Strong Convexity $\exists m > 0$ s.t. $\nabla^2 f(x) \succeq mI$ for all $x \in S$.

e^{-x} is strictly convex but not strongly convex.

Big Picture Results for Gradient Descent from last lecture, assuming f strongly convex.

Exact Line Search: At least as good as $t = \frac{1}{M}$,

which gives

$$f(x^{(k)}) - p^* \leq \left(1 - \frac{1}{K}\right)^k \left(f(x^{(0)}) - p^*\right)$$

where $K = M/m$

$$M = \sup_{x \in S} \lambda_{max}(\nabla^2 f(x))$$

$$m = \inf_{x \in S} \lambda_{min}(\nabla^2 f(x))$$

$\Rightarrow$ # iters for accy ε is $O\left(\frac{1}{\log \varepsilon}\right)$

BUT don't know M, m in practice.

Hence: BACKTRACKING LINE SEARCH

$\alpha \in (0, 1/2)$ ARMIJO COND

$\beta \in (\alpha, 1)$ BACKTRACK PARAM

When $\beta = \frac{1}{2}$, $M \geq 1/2$, get

$$f(x^{(k)}) - p^* \leq \left(1 - \frac{\alpha}{K}\right)^k \left(f(x^{(0)}) - p^*\right).$$

Gradient Descent, cont'd.

Nesterov (Thm 2.1.15) also gives another complexity result for

$$t = \frac{2}{m+M}$$

namely

$$\|x^{(k)} - x^*\| \leq \left(\frac{1-1/K}{1+1/K}\right)^k \|x^{(0)} - x^*\|$$

which leads to

$$f^{(k)} - p^* \leq K \cdot \left(\frac{1-1/K}{1+1/K}\right)^{2k} \left(f^{(0)} - p^*\right)$$

where x^* = minimizer, $p^* = f(x^*)$.

Both BV and Nesterov also give results for the non-(strongly convex) case — these are much weaker.

NMI

Newton's Method.

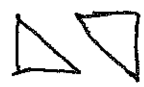
Δx is solution of $\nabla^2 f(x^{(k)})\,\Delta x = \overbrace{-\nabla f(x^{(k)})}^{-g}$.

Motivation: minimize "quadratic model"

$$\varphi(v) \underset{\substack{\uparrow \\ [f(x^{(k)})+]}}{=} \nabla f(x^{(k)})^T v + \frac{1}{2} v^T \nabla^2 f(x^{(k)})\, v.$$

To solve equation, use CHOLESKY FACTORIZATION

$$\nabla^2 f(x^{(k)}) = L L^T$$



$$L \underbrace{L^T \Delta x}_{y} = -g.$$

1) forward solve for y
2) back solve for Δx.

Cost: $\frac{1}{6} n^3$ adds + mults.

Use same backtracking line search.

(Newton used this for finding zeros of polynomials, not minimization; particularly root of $p(\lambda) = \lambda^2 - a$ i.e. square roots).

NM2

Convergence Analysis of Newton's Method.

As before, suppose that $S = \{x : f(x) \le f(x_0)\}$ is compact and $MI \succeq \nabla^2 f(x) \succeq mI$ on S, $m > 0$

Now also need

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L\|x - y\| \quad \forall x, y \in S$$

i.e. $\nabla^2 f$ is Lipschitz.

Turns out (BV §9.5) that $\exists \eta > 0, \gamma > 0$ s.t.

If $\|\nabla f(x^{(k)})\| \ge \eta$, then backtracking line sch returns $t_k \ge \frac{\beta m}{M}$ with

$$f(x^{(k+1)}) - f(x^{(k)}) \le -\gamma \qquad (\dagger)$$

while if $\|\nabla f(x^{(k)})\| < \eta$, B.T.L.S. returns $t_k = 1$ with

$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\| \le \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|\right)^2 \qquad (*)$$

NOTE

"QUADRATIC CONVERGENCE".

Consequence: If $\eta \le \frac{m^2}{L}$ and, for some K, $\|\nabla f(x^{(K)})\| \le \eta$

then

NM3

$$\|\nabla f(x^{(K+1)})\| \le \frac{L}{2m^2}\eta^2 \le \frac{1}{2}\eta$$

so this applies recursively and hence (*) holds for all $\ell \ge K$, so

$$\frac{L}{2m^2}\|\nabla f(x^{(\ell)})\| \le \left(\frac{L}{2m^2}\|\nabla f(x^{(K)})\|\right)^{2^{\ell-K}}$$

$$\le \left(\frac{1}{2}\right)^{2^{\ell-K}} \qquad (\ddagger)$$

	$\ell=K$	$\ell=K+1$	$\ell=K+2$	$\ell=K+3$	
RHS	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{256}$	quad. conv.

But what are η, γ? Turns out (BV p. 489–491) that using BTLS with Newton step, we always have $t_k \ge \frac{\beta m}{M}$
and that consequently

$$f(x^{(k+1)}) - f(x^{(k)}) \le \underbrace{-\alpha\beta\eta^2\frac{m}{M^2}}_{\gamma}$$

so (†) holds if we set γ.

It also turns out that if $\eta \le 3(1-2\alpha)\frac{m^2}{L}$

then $t_k = 1$, i.e. $f(x^{(k)} + \underbrace{\Delta x_{nt}}_{\text{Newton step}})$ satisfies the "sufficient decrease" condition in the BTLS.

We will now show that (*) holds as a consequence.
(QUAD CONTR.)

Proof of (*) (quadratic contraction) assuming $t_k = 1$.

$$\|\nabla f(x^{(k)} + \Delta x_{NT})\| = \| \nabla f(x^{(k)} + \Delta x_{NT}) \underbrace{- \nabla f(x^{(k)}) - \nabla^2 f(x^{(k)}) \Delta x_{NT}}_{\text{ZERO BY DEF.}} \|$$

$$\nabla f(x^{(k)} + \Delta x_{NT}) - \nabla f(x^{(k)}) = \int_0^1 \nabla^2 f(x^{(k)} + s\Delta x_{NT}) \Delta x_{NT} \, ds$$

fund thm calc.

$$= \left\| \int_0^1 \left(\nabla^2 f(x^{(k)} + s\Delta x_{NT}) - \nabla^2 f(x^{(k)}) \right) \Delta x_{NT} \, ds \right\|$$

$$\le \int_0^1 L \, \|s \Delta x_{NT}\| \, \|\Delta x_{NT}\| \, ds$$

$$= \frac{L}{2} \|\Delta x_{NT}\|^2$$

$$= \frac{L}{2} \left\| \left(\nabla^2 f(x^{(k)}) \right)^{-1} \nabla f(x^{(k)}) \right\|^2$$

$$\le \frac{L}{2m^2} \|\nabla f(x^{(k)})\|^2 \qquad \equiv (*).$$

Note: for this to apply recursively, also need $\eta \le \frac{m^2}{L}$ as explained on NM2 (bottom). So, need

$$\eta = \min(1, 3(1-2\alpha)) \frac{m^2}{L}.$$

NM5

Total # iterations

Initial Phase with $\|\nabla f(x)\| \geq \eta$

$$\#\text{steps} \leq \frac{f(x^{(0)}) - p^*}{\gamma} \text{ immediately from } (\dagger).$$

Quadratically convergent phase with $\|\nabla f(x)\| < \eta$:

$$f(x^{(l)}) - p^* \leq \frac{1}{2m} \|\nabla f(x^{(l)})\|^2 \leq \quad \leftarrow \text{(from 1')}$$

(square both sides of ($\ddagger$)) (BV p 467, line 5)

$$\leq \frac{1}{2m} \frac{4m^4}{L^2} \left(\frac{1}{2}\right)^{2^{l-K+1}}$$

$$= \frac{2m^3}{L^2} \left(\frac{1}{2}\right)^{2^{l-K+1}}$$

If want RHS $\leq \varepsilon$, or LHS $\leq \varepsilon$, need

$$\left(\frac{1}{2}\right)^{2^{l-K+1}} \leq \frac{\varepsilon L^2}{2m^3}$$

$$2^{2^{l-K+1}} \geq \frac{\varepsilon_0}{\varepsilon} \quad \text{where } \varepsilon_0 = \frac{2m^3}{L^2}$$

$$2^{l-K+1} \geq \log_2 \frac{\varepsilon_0}{\varepsilon}$$

$$\#\text{steps} = l-K+1 \geq \log_2 \log_2 \frac{\varepsilon_0}{\varepsilon}$$

$$\text{e.g. } \frac{\varepsilon_0}{\varepsilon} = 10^{15}$$

$$\log_2 10^{15} \approx \log_2 2^{50} = 50$$

$$\log_2 50 \leq 6.$$

#ACCURATE DIGITS DOUBLES EVERY ITERATION.

Very few steps once quadratic convergence starts.

LCB 1

Lower Complexity Bounds (Nesterov Sec 2.1.4)

Assume as before that f is strongly convex & C^2 with

$$mI \preceq \nabla^2 f(x) \preceq MI \quad \text{for all } x \in S.$$

m: μ in Nesterov; M: L in Nesterov.

Assume that at each point $x^{(k)}$, a "first-order oracle" or "black box" computes $f(x^{(k)})$ and $\nabla f(x^{(k)})$. Assume also that for $k = 1, 2, \ldots$

(ASSUMPTION A) $\quad x_k \in \text{Linear Span}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\}$.

For simplicity, assume $\operatorname{dom} f = \mathbb{R}^\infty \equiv \ell_2 =$

$$\left\{ x = (x_i)_{i=1}^{\infty} : \|x\|^2 = \sum_{i=1}^{\infty} x_i^2 < \infty \right\}.$$

Now we define a "difficult" function F by

$$F(x) = \frac{M-m}{8}\left\{ (x_1)^2 + \sum_{i=1}^{\infty}(x_i - x_{i+1})^2 - 2x_1 \right\} + \frac{m}{2}\|x\|^2.$$

We have

$$\frac{\partial F}{\partial x_1} = \frac{M-m}{8}\left(2x_1 + 2(x_1 - x_2) - 2\right) + m x_1$$

$$j > 1: \quad \frac{\partial F}{\partial x_j} = \frac{M-m}{8}\left(2(x_j - x_{j+1}) - 2(x_{j-1} - x_j)\right) + m x_j$$

$$= \frac{M-m}{8}\left(4x_j - 2x_{j+1} - 2x_{j-1}\right) + m x_j$$

LCB2

so
$$\nabla^2 F(x) = \frac{M-m}{4} \underbrace{\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & \ddots \\ & & \ddots & \ddots \end{bmatrix}}_{\preceq 4I,\ \succeq 0} + mI$$

with ~~$\nabla F(x) =$~~

$$\left.\begin{aligned} \lambda_{max}\left(\nabla^2 F(x)\right) &\leq M \\ \lambda_{min}\left(\nabla^2 F(x)\right) &\geq m \end{aligned}\right\} \text{as required.}$$

and

$$\nabla F(x) = \left(\frac{M-m}{4} T + mI\right) x - \frac{M-m}{4} e_1 \leftarrow \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

The solution x^* is given by $\nabla F(x^*) = 0$:

$$\frac{M-m}{4} (2x_1 - x_2) + m x_1 = \frac{M-m}{4}$$

$$2x_1 - x_2 + \frac{4}{M-m} m x_1 = 1$$

$$x_2 - \frac{2(M-m) + 4m}{M-m} x_1 + 1 = 0$$

$$x_2 - 2\frac{M+m}{M-m} x_1 + 1 = 0$$

and, for $j = 2, 3, \ldots$

$$x_{j+1} - 2\frac{M+m}{M-m} x_j + x_{j-1} = 0$$

This difference equation can be solved by plugging in $x_j = q^j$ and solving for q:

LCB3

$$q^{j+1} - 2\frac{M+m}{M-m}q^{j} + q^{j-1} = 0$$

$$q^{2} - 2\frac{M+m}{M-m}q + 1 = 0$$

Claim: roots are $\dfrac{M+m \pm 2\sqrt{M}\sqrt{m}}{M-m}.$

check: sum of roots is then $2\dfrac{M+m}{M-m}$ ✓

product of roots is $\dfrac{(M+m)^2 - 4Mm}{(M-m)^2} = 1$ ✓

<u>smaller</u> root is

$$q = \frac{M+m-2\sqrt{M}\sqrt{m}}{M-m} = \frac{(\sqrt{M}-\sqrt{m})^2}{(\sqrt{M}-\sqrt{m})(\sqrt{M}+\sqrt{m})}$$

$$= \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}$$

$$= \frac{1-\sqrt{1/K}}{1+\sqrt{1/K}} \qquad \text{where } K = \sqrt{\frac{M}{m}}$$

NOTE THE SQRT.

LCB4

<u>We get</u> <u>Theorem</u> For any $x^{(0)} \in \mathbb{R}^\infty$ and any $m > 0$, $M > m$, $\exists$ function F (quadratic) with $mI \le \nabla^2 F \le MI$ with (under ASSUMPTION A)

$$\|x^{(k)} - x^*\|^2 \geq \left(\frac{1-\sqrt{1/K}}{1+\sqrt{1/K}}\right)^{2k} \|x^{(0)} - x^*\|^2$$

where x^* minimizes F and $K = M/m$, and hence

$$f(x^{(k)}) - f^* \geq \frac{m}{2}(\text{ same })^{2k} \|x^{(0)} - x^*\|^2$$

<u>Pf</u> WLOG take $x^{(0)} = 0$. Then we use F as defined already, getting:

$$\|x^{(0)} - x^*\|^2 = \sum_{i=1}^{\infty} (x^*)_i^2 = \sum_{i=1}^{\infty} q^{2i} = \frac{q^2}{1-q^2}$$

Since T is tridiagonal and $\nabla f(x^{(0)}) = e_1$, we can show by induction that $x_j \in \mathrm{span}(e_1, e_2, \ldots, e_j)$, so

$$\|x^{(j)} - x^*\|^2 \geq \sum_{i=j+1}^{\infty} (x^*)_j^2 = \sum_{i=j+1}^{\infty} q^{2i}$$

$$= \frac{q^{2(j+1)}}{1-q^2} = q^{2j} \|x^{(0)} - x^*\|^2$$

with $q = \dfrac{1-\sqrt{1/K}}{1+\sqrt{1/K}}$. (Last inequality follows from our original (1), just Taylor's Thm.)
(BV (9.8))

LCB5

In comparison, the gradient method with $t \equiv \frac{1}{M}$ gave us

$$f(x^{(k)}) - f^* \leq \left(1 - \frac{1}{K}\right)^k (f(x^{(0)}) - f^*)$$

Compare $$\frac{1 - \sqrt{1/K}}{1 + \sqrt{1/K}} \approx \left(1 - \sqrt{1/K}\right)^2 \approx 0.998$$

if $K = 10^6$, while $1 - \frac{1}{K} \approx 0.999999$.

So lower bound indicates we may be able to do much better. The rest of Nesterov's Chapter 2 derives "optimal gradient" methods, but the argument is very complicated! In the end, the simplest is:

NESTEROV OPTIMAL GRADIENT ALGORITHM (p. 81)

Choose $y^{(0)} = x^{(0)} \in \mathbb{R}^n$

For $k = 0, 1, 2, \ldots$

$$\left[\begin{array}{l} \text{set } x_{k+1} = y_k - \frac{1}{M} \nabla f(y_k) \\ \text{set } y_{k+1} = x_{k+1} + q\,(x_{k+1} - x_k) \end{array}\right.$$

where $q = (1 - \sqrt{1/K}) / (1 + \sqrt{1/K})$