Back to GD:

We've seen that

$$\left.\begin{array}{l} \|x^{(0)} - x^*\| \le R \\ \|\nabla f(x)\| \le L \\ f \text{ convex} \end{array}\right\} \Rightarrow f\left(\frac{1}{t}\sum_{s=0}^{t-1} x^{(s)}\right) - f(x^*) \le \frac{RL}{\sqrt{t}}$$

when $\mu = \frac{R}{L\sqrt{t}}$

So, for an error $\le \varepsilon$, we require $\approx \frac{1}{\varepsilon^2}$ iterations
Can we do better if we assume more about f?

Defn: We say $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth if its gradient is L-Lipschitz
i.e. if

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$$

Theorem: If $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth & convex & $0 \le \mu \le \frac{1}{L}$ then the GD iterations

$$x^{(t+1)} = x^{(t)} - \mu \nabla f(x^{(t)})$$

satisfy

$$f(x^{(t)}) - f(x^*) \le \frac{1}{2t\mu}\|x^{(0)} - x^*\|$$

* We won't prove this theorem, but a key element in its proof is the fact that for L-smooth convex functions, we have that

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|_2^2$$

* With L-smoothness, need $\approx \frac{1}{\varepsilon}$ iterations to get ε error.

* Can modify GD to get projected GD for L-smooth function with same convergence rate when solving

$$\min_{x \in \Omega} f(x) \qquad (\text{with } \Omega \subset \mathbb{R}^n, \text{ convex})$$

Picking μ in Practice:

* The convergence theorems we've seen so far require knowing the Lipschitz constant or smoothness constant associated with $f(x)$ (in order to set μ).

* This is not always possible in practice!

* Idea: Use the "best" $\mu^{(t)}$ at every iteration

$$x^{(t+1)} = x^{(t)} - \mu^{(t)} \nabla f(x^{(t)})$$

pick $\mu^{(t)}$ to minimize

$$\min_{\mu(t)} f(\underbrace{x^{(t)}}_{\text{known}} - \underbrace{\mu^{(t)}}_{\text{unknown}} \underbrace{\nabla f(x^{(t)})}_{\text{known}})$$

[the variable we're optimizing here is $\mu(t)$.

Often, solving this exactly is hard so we settle for an approximate solution.

* Possible Solution: set μ using a Backtracking Line Search at every iteration t.

Idea: for any descent direction p

$$f(x) - \nabla f(x)^T p \leq f(x-p) \leq f(x) - \gamma \nabla f(x)^T p$$

($*$)

f is convex

p is a descent dir.

So there is a small constant $\gamma > 0$, that make this ineq. true

Take $p = \mu \nabla f(x)$ (as in GD) and plug it into ($*$).

So we expect that if μ is small enough

$$f(x - \mu \nabla f(x)) \leq f(x) - \gamma \mu \|\nabla f(x)\|_2^2$$

* Idea: Fix γ (say 0.5)

Start with (e.g.) $\mu = 1$

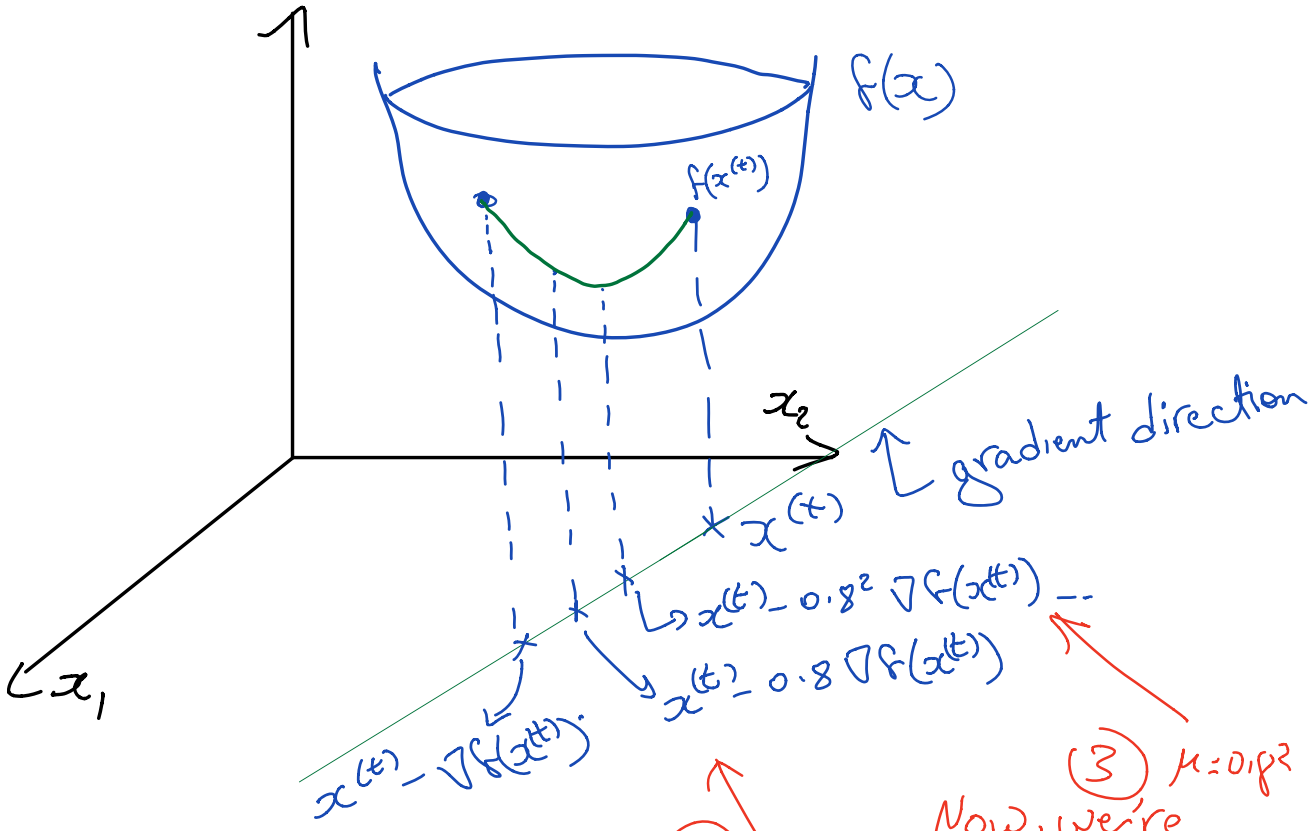
decrease μ iteratively until

$f(x^{(t+1)}) = f(x^{(t)} - \mu^{(t)} \nabla f(x^{(t)}))$

is smaller than $f(x^{(t)}) - \gamma \mu \|\nabla f(x^{(t)})\|^2$

e.g.: $\mu = 1 \longrightarrow \mu = 0.8 \longrightarrow \mu = 0.8^2 \cdots$

"Example by picture"



① $\mu = 1$: $f(x^{(t)} - \mu \nabla f(x^{(t)}))$ is too big (bigger than $f(x^{(t)}) - \mu\alpha \|\nabla f(x^{(t)})\|^2$)

② So we try $\mu = 0.8$ (still too big)

③ $\mu = 0.8^2$ Now, we're good

More Precisely:

Backtracking Line Search

Pick $\beta < 1$, $\gamma < 1$

At each GD step t:

(1) Set $v = -\nabla f(x^{(t)})$

(2) Set $\mu = 1$

(3) If

$$f(x^{(t)} - \mu \nabla f(x^{(t)})) \leq f(x^{(t)}) - \mu\gamma \|\nabla f(x^{(t)})\|^2$$

Then

keep μ

Else

Set $\mu \leftarrow \beta\mu$ & Repeat (3)

Remark: The condition

$$f(x^{(t)} - \mu \nabla f(x^{(t)})) \leq f(x^{(t)}) - \mu\gamma \|\nabla f(x^{(t)})\|^2$$

is known as the Armijo Condition and guarantees that the function value decreases by some non-zero amount.

Example: Consider the function

$f: \mathbb{R}^2 \to \mathbb{R}$ with

$$f(x_1, x_2) = (x_1 - 1)^4 + (x_1 + x_2 - 1)^2$$

(which has minimizer $(1,0)$)

$$\nabla f = \left(4(x_1-1)^3 + 2(x_1+x_2-1), 2(x_1+x_2-1)\right)$$

Suppose $x^{(0)} = (0,0)$

$$x^{(1)} = x^{(0)} - \mu \nabla f(x^{(0)}) = (0,0) - \mu(-6,-2)$$

Want to pick μ to minimize $f((0,0) - \mu(-6,-2))$

$$f((0,0) - \mu(-6,-2)) = f(6\mu, 2\mu) = (6\mu - 1)^4 + (8\mu - 1)^2$$

This is a non-linear eq'n in μ that may be hard to minimize

Then let's run Backtracking Linesearch with $\gamma = 0.5$, $\beta = 0.8$

$$x^{(1)} = x^{(0)} - \mu \nabla f(x^{(0)})$$
$$= (0,0) - \mu(-6,-2)$$

So we try. $\mu = 1$:

$f(x^{(1)}) = 674 > f(x^{(0)}) - \mu\gamma \|\nabla f(x^{(0)})\|^2$

(2, 1, 0.5, 40; -18)

$x^{(0)} - \mu \nabla f(x^{(0)})$

$\mu = 0.8$:

$f(x^{(1)}) = 237.6736 > f(x^{(0)}) - \mu\gamma \|\nabla f(x^{(0)})\|^2$ (-14)

$\mu = 0.8^2$ (same issue)

Finally $\mu = 0.8^{11}$ gives

$f(x^{(1)}) = 0.1530 \leq f(x^{(0)}) - 0.5 \times 0.8^{11} \|\nabla f(x^{(0)})\|^2$

So we choose $\mu = 0.8^{11}$ and set $x^{(1)} = (0,0) - 0.8^{11} \nabla f(0,0)$

$= (0.5154, 0.1718)$

* Then we repeat the process for $x^{(2)}, \ldots x^{(1000)} \ldots$

$\hookrightarrow (0.979, 0.0211)$

Fact/Theorem: For an L-smooth convex function, with $\mu^{(t)}$ set by backtracking line search as above, we have

$$f(x^{(t)}) - f(x^*) \leq \frac{1}{2t \min_{s=1,\dots,t} (\mu^{(s)})}$$

and $\min_{s=1,\dots,t} \mu^{(s)} \geq \min(1, \beta/L)$

↳ so we get $f(x^{(t)}) - f(x^*) \leq \frac{L}{2t\beta}$

* More can be said when we assume more about the function (e.g. strong convexity).
* More can be said about line search methods.
* We may return to these topics later if time permits.

Newton's Method

* So far, we have used GD, which was derived from the 1st order Taylor approximation

$$f(x) \approx f(x^{(t)}) + \nabla f(x^{(t)})(x - x^{(t)})$$

↳ want this small $\Rightarrow$ want this as neg. as possible

$\Rightarrow$ pick $x^{(t+1)} = x^{(t)} - \mu \nabla f(x^{(t)})$

to get $f(x^{(t+1)}) \approx f(x^{(t)}) - \mu \|\nabla f(x^{(t)})\|^2$

* If, instead, we use a 2nd order Taylor approximation, we obtain Newton's method

$$f(x) \approx f(x^{(t)}) + \nabla f(x^{(t)})^T (x - x^{(t)}) + \frac{1}{2}(x - x^{(t)})^T \nabla^2 f(x^{(t)})(x - x^{(t)})$$

At a minimum, we expect that $\nabla f = 0$ (because the right hand side is convex)

so taking ∇ on both sides

$$\nabla f(x) \approx 0 + \nabla f(x^{(t)}) + \nabla^2 f(x^{(t)})(x - x^{(t)})$$

$$\approx \Rightarrow \quad x - x^{(t)} \approx -\left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})$$

at the minimizer, so we set

$$\boxed{x^{(t+1)} = x^{(t)} - \left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})}$$

Example: Let $f : \mathbb{R}^+ \to \mathbb{R}$ be given by $f(x) = x - \ln(x)$

Then $\nabla f(x) = f'(x) = 1 - \frac{1}{x}$

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

Newton's method starting at $x^{(0)} = 0.5$ generates the iterates

$$x^{(t+1)} = x^{(t)} - \left[f''(x^{(t)})\right]^{-1} f'(x^{(t)})$$

$$= x^{(t)} - \left[x^{(t)}\right]^2 \left(1 - \frac{1}{x^{(t)}}\right)$$

$$= x^{(t)} - \left[x^{(t)}\right]^2 + x^{(t)} = 2x^{(t)} - \left(x^{(t)}\right)^2$$

$$\Rightarrow x^{(1)} = 2x^{(0)} - \left[x^{(0)}\right]^2 = 1 - 0.25 = 0.75$$

$x^{(2)} = \ldots = 0.9375$

$x^{(3)} = \ldots = 0.9961$

$x^{(4)} = \ldots = 0.9998$

(Note that the optimum is $x^* = I$)

If this seems fast, it is not a coincidence

Def'n: for a matrix M:

$$\|M\| = \max_{x \neq 0} \frac{\|Mx\|}{\|x\|}$$

$\|Mx\|$ → length of Mx

$\|x\|$ → length of x

↑ $\|M\|$ measures how much the matrix M can "stretch" vectors

(Consequence of the def'n:

$\forall z: \quad \|Mz\| \leq \|M\| \, \|z\|$

↑ length of Mz ↑ norm of M ↑ "length" of z)

Theorem: Let f be twice cont. differentiable & suppose that x^* has $\nabla f(x^*) = 0$.

Suppose further that

$$\begin{cases} \| (\nabla^2 f(x^*))^{-1} \| \leq \frac{1}{h} & \text{for some } h > 0 \\ \| \nabla^2 f(x) - \nabla^2 f(x^*) \| \leq L \| x - x^* \| \\ \quad \text{for all } x \end{cases}$$

Then, if $\| x^{(0)} - x^* \| \leq \frac{2h}{3L}$

($x^{(0)}$: starting pt.)

& $x^{(t+1)} = x^{(t)} - \left[\nabla^2 f(x^{(t)}) \right]^{-1} \nabla f(x^{(t)})$

($x^{(t+1)}$: next pt.)

we have:

$$\begin{cases} \| x^{(t)} - x^* \| \leq \frac{2h}{3L} \quad \forall t \\ \| x^{(t+1)} - x^* \| \leq \frac{3L}{2h} \| x^{(t)} - x^* \|^2 \\ \qquad \forall t \end{cases}$$

Loose Interpretation: If we start close enough to a local minimum, and the function is nice, we converge quickly to the minimum.

Example

Want to minimize
$f(x) = x_1^4 + 2x_1^2x_2^2 + x_2^4$
using Newton's Method

$\Rightarrow$ need $\nabla f(x)$ & $\nabla^2 f(x)$

$$\nabla f(x) = \begin{pmatrix} 4x_1^3 + 4x_1x_2^2 \\ 4x_1^2x_2 + 4x_2^3 \end{pmatrix}$$

$$\nabla^2 f(x) = \begin{pmatrix} 12x_1^2 + 4x_2^2 & 8x_1x_2 \\ 8x_1x_2 & 4x_1^2 + 12x_2^2 \end{pmatrix}$$

Suppose we start at $x^{(0)} = (1,1)$

Then $x^{(1)} = x^{(0)} - \left[\nabla^2 f(x^{(0)})\right]^{-1} \nabla f(x^{(0)})$

$$= \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{bmatrix} 16 & 8 \\ 8 & 16 \end{bmatrix}^{-1} \begin{pmatrix} 8 \\ 8 \end{pmatrix}$$

$$= \begin{pmatrix} 2/3 \\ 2/3 \end{pmatrix}$$

Continuing this way

$$x^{(t)} = \begin{pmatrix} 2/3 \\ 2/3 \end{pmatrix}^t \xrightarrow[t \to \infty]{} 0$$

exponentially fast in t.

Example

$f(x) = \frac{x^4}{4} - x^2 + 2x + 1$

start at $x^{(0)} = 0$

$\nabla f(x) = x^3 - 2x + 2 \qquad \nabla^2 f(x) = 3x^2 - 2$

$$\Rightarrow x^{(1)} = x^{(0)} - \left[\nabla^2 f(x^{(0)})\right]^{-1}\left(\nabla f(x^{(0)})\right)$$

$$= 0 - (-2)^{-1}\begin{pmatrix} 2 \end{pmatrix}$$

$$= 1$$

$$x^{(2)} = x^{(1)} - \left[\nabla^2 f(x^{(1)})\right]^{-1}\nabla f(x^{(1)})$$

$$= 1 - 1^{-1}\begin{pmatrix} 1 \end{pmatrix}$$

$$= 0 = x^{(0)}$$

back to $x^{(0)}$! So we entered a cycle!

So, Newton's method need not always converge.

Why does this not contradict the theorem?

Some Remarks on Newton's Method

(1) We can modify Newton's method to include a step size $0 < \mu < 1$

$$x^{(t+1)} = x^{(t)} - \mu^{(t)} \left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})$$

↑ can choose fixed
- can use backtracking line search.

(2) Finding the inverse of the Hessian can be very expensive if n is large.

Instead, notice that

$$x^{(t+1)} = x^{(t)} - \left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})$$

$$\Rightarrow \left[\nabla^2 f(x^{(t)})\right] \left[x^{(t+1)} - x^{(t)}\right] = -\nabla f(x^{(t)})$$

known — solve for this — known — known

So we can use linear algebra techniques to solve for $x^{(t+1)}$ from this system.

(3) Newton's method, when it converges, tends to do so much faster than GD. (compare the convergence theorems).

Quasi-Newton Methods

(very briefly)

Recall: GD has the interpretation

$$f(x) \approx f(x^{(t)}) + \nabla f(x^{(t)})^T (x - x^{(t)}) + \frac{\theta_t}{2} \|x - x^{(t)}\|^2$$

minimizing the right hand side gives

$$x^{(t+1)} = x^{(t)} - \frac{1}{\theta_t} \nabla f(x^{(t)})$$

Meanwhile Newton's Method

$$f(x) \approx f(x^{(t)}) + \nabla f(x^{(t)})^T (x - x^{(t)}) + \frac{1}{2}(x - x^{(t)})^T \nabla^2 f(x^{(t)})(x - x^{(t)})$$

minimizing the right hand side gives

$$x^{(t+1)} = x^{(t)} - \left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})$$

So, we can think of GD as approximating the Hessian with

$$\frac{1}{\sigma_t} I$$

↑ Identity matrix.

Quasi-Newton Methods approximate the Hessian with some other matrix B_t which changes from iteration to iteration, so that

$$x^{(t+1)} = x^{(t)} - \mu^{(t)} B_t^{-1} \nabla^2 f(x^{(t)})$$

There are several such methods (with different choices for B_t)

We won't cover them here, but examples are:

BFGS method

Broyden

⋮