Convex Functions, part 3: Lipschitz gradient, etc.

Sunday, January 31, 2021 3:49 PM

mostly details not in BV'04

Recall a function $F: \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous if $\exists\, L>0$ s.t. $\forall x,y \in \mathrm{dom}(F)$,

$\|F(x)-F(y)\| \le L\cdot\|x-y\|$.

If F' exists, then $|||F'||| \le L \implies F$ is Lipschitz continuous

appropriate operator norm, usually $|\cdot|$ if 1D or spectral norm

What do we mean by this?

The Jacobian, where if $F(x) = \begin{bmatrix} F_1(x) \\ \vdots \\ F_m(x)\end{bmatrix}$, $F_i: \mathbb{R}^n \to \mathbb{R}$,

$J_{ij} = \dfrac{\partial F_i(x)}{\partial x_j}$ (or $\dfrac{\partial F_j(x)}{\partial x_i}$, I can never remember, and conventions aren't consistent anyhow)

$m\times n$ (or $n \times m$) matrix.

In optimization, "Jacobian" is often confusing, since it's unclear what "F" is.

Ex: $f: \mathbb{R}^n \to \mathbb{R}$, $f(x)$ a scalar

$\nabla f: \mathbb{R}^n \to \mathbb{R}^n$, "gradient", $\nabla f(x)$ a vector, "operates" on directions d like $\langle \nabla f(x), d\rangle$ (so a linear operator)

$\nabla^2 f: \mathbb{R}^n \to \mathbb{R}^{n\times n}$, "Hessian", $\nabla^2 f(x)$ a matrix, "operates" on d (symmetric matrix) like $\langle d, \nabla^2 f(x)\, d\rangle$

$(\nabla f(x))_i = \dfrac{\partial f}{\partial x_i}$

$(\nabla^2 f(x))_{ij} = \dfrac{\partial^2 f}{\partial x_i \partial x_j}$

... so Jacobian of $F=f$ is the gradient (transposed)

... but Jacobian of $F = \nabla f$ is the Hessian

Fact Suppose $f \in C^2(\mathcal{U})$ for some open set $\mathcal{U} \subseteq \mathbb{R}^n$, and f convex, then

$\big(\nabla f$ is L-Lipschitz cts. on $\mathcal{U}\big)$ iff $\big(\forall x \in \mathcal{U},\ \nabla^2 f(x) \preceq L I\big)$

i.e., all eigenvalues $(\nabla^2 f(x)) \le L$

$\implies |||\nabla^2 f(x)||| \le L$

Fact Suppose $f \in C^2(\mathcal{U})$ for some open set $\mathcal{U} \subseteq \mathbb{R}^n$, then

$\big(f$ is μ-strongly convex on $\mathcal{U}\big)$ iff $\big(\forall x \in \mathcal{U},\ \mu I \preceq \nabla^2 f(x)\big)$

(w/ respect to $\|\cdot\|_2$)

(need $\mu > 0$. If $\mu = 0$ this is plain old convexity)

So... one of our common assumptions will be $\boxed{\nabla f \text{ is } L \text{ Lipschitz}}$ $(\nabla^2 f \preceq L I)$

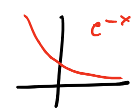
and, a bit less often, also assume strong convexity $(\nabla^2 f \succeq \mu I)$

often called "strong smoothness" recently

$\boxed{Q1}$ Is $f(x) = e^{-x}$ convex, strictly convex, strongly convex ?
Is f' Lipschitz ?

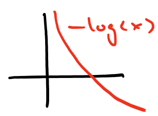
$\boxed{Q2}$ Is $f(x) = \begin{cases} -\log(x) & x > 0 \\ +\infty & x \le 0 \end{cases}$ convex.... ?
f' Lipschitz ?

$\boxed{A1}$ e^{-x} on $\mathbb{R}$ is strictly convex (hence convex) but not strongly cvx
$(f'(x) = e^{-x} \to 0$ as $x \to \infty)$. It is strongly convex on the domain $(-\infty, R]$ for any $R < \infty$.



Similarly, f' isn't Lipschitz on $\mathbb{R}$ but it is Lipschitz on $[-R, \infty)$ $\forall R < \infty$

$\boxed{A2}$ $-\log(x)$ is convex and strictly convex



since $f' = \frac{-1}{x}$ is monotone

... but not strongly convex unless we look at $(-\infty, R]$ again.

For f' Lipschitz, it isn't on $(0, \infty)$ but it is on $[\delta, \infty)$ $\forall \delta > 0$. } Can create problems for some algorithms if they converge to $x = 0$

Ex $f(x) = \frac{1}{2}\|x\|_2^2$, $\nabla f(x) = x$, $\nabla^2 f(x) = I$
$\Rightarrow L = 1$, $\mu = 1$. Only function w/ this property.

NICEST FUNCTION EVER :

Calculus : $f(x) = e^x$

Statistics : $f(x) = e^{-x^2/2}$ or multivar. version

Optimization : $f(x) = \frac{1}{2} x^2$ (= negative log-likelihood of Gaussian!)

Def The condition # of f is $\kappa_f = L/\mu$. $\kappa_f \approx 1$ good
$\kappa_f \approx \infty$ bad

Why these assumptions ?

Taylor's Thm: $f(y) = f(x) + f'(x)\cdot(y-x) + \frac{1}{2} f''(\xi)(y-x)^2$

for some $\xi \in [x, y]$ (or in $[y, x]$). Similar in higher-dim.

if $f''(\xi) \le L \ \forall \xi$, then

$f(y) = \ldots \le f(x) + f'(x)(y-x) + \frac{1}{2} L (y-x)^2$.

so...

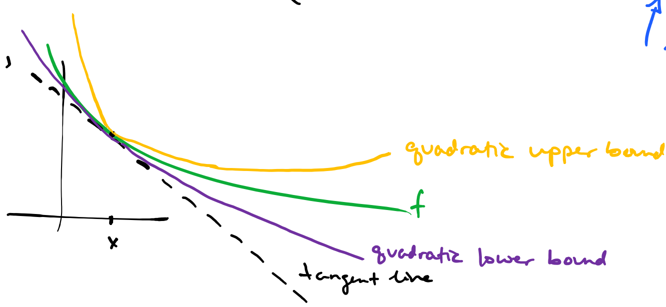
Theorem If ∇f is L-Lipschitz and f is μ-strongly convex (and take $\mu = 0$ if just convex)

then $\forall x, y \in \mathrm{dom}(f)$,

$$\mu/2 \|y-x\|^2 \le f(y) - \left(f(x) + \langle \nabla f(x), y-x \rangle \right) \le L/2 \|y-x\|^2$$

This inequality is sometimes called "The Descent Lemma"

Graphically,



Usually f is complicated, but now we can "sandwich" it between a quadratic upper bound and a quadratic lower bound (if strongly cvx, $\mu > 0$) or a linear lower bound (if just cvx, $\mu = 0$)

and quadratics are easy to work with, eg. easy to minimize in closed form, etc.

More properties

eg. f convex $\Rightarrow$ ∇f monotone, meaning $\langle x-y, \nabla f(x) - \nabla f(y) \rangle \ge 0$

These can be strengthened w/ our $\mu > 0$ and L assumptions

See github class website $\boxed{\text{Handouts / StrongConvexityLipschitz.pdf}}$