26. Convex Learning Problems

Monday, April 1, 2024 8:44 AM

ch. 12 in [SS]

multivariate calc review:
∇f is gradient
$\nabla^2 f$ is Hessian
'Jacobian' can be ambiguous

Ch. 13 will connect to ML, this chapter is quick summary of optim.
See [Boyd + Vandenberghe] book or Appendix in [Mohri et al.]

Convex Functions

$f: \mathbb{R}^d \to \mathbb{R}$ is convex if $\forall x, y \in$ domain,

1) $\forall \alpha \in [0,1]$, $f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$
2) ... or, if $\text{epi}(f)$ is a convex set in $\mathbb{R}^{d+1}$
3) ... or, if all 1D restrictions of f are (1D) convex functions
   i.e. $\forall x, \forall y$, $\varphi: t \mapsto f(x + ty)$ is convex

4) ... or, if ∇f exists, ∇f is monotone (in 1D, means $x \geq y \Rightarrow f'(x) \geq f'(y)$)
   i.e. $\forall x, y \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$
5) ... or, if $\nabla^2 f$ exists, $\nabla^2 f$ is positive semidefinite (in 1D, means $f''(x) \geq 0$)
   i.e. $\forall x, y \quad y^T \nabla^2 f(x)\, y \geq 0$ i.e. all eigenvalues of Hessian are non-negative

How to prove a function is convex?

- building blocks of simpler functions:
  f, g convex $\Rightarrow$ $\alpha f + \beta g$ is convex ($\forall \alpha, \beta \geq 0$)
- use #3 or #5 definitions above ... use #1 only as last resort!
- compositions: f, g convex, usually $f \circ g$ isn't convex
  (see B+V for sufficient conditions)
  but... f convex, g affine $\Rightarrow$ $f \circ g$ is convex
  ex: prove $f(x) = \frac{1}{2}\|Ax - b\|^2$ is convex
- other calculus, eg. A, B convex sets $\Rightarrow$ $A \cap B$ a convex set
  f, g convex fcn $\Rightarrow$ $x \mapsto \sup(f(x), g(x))$ is convex

Subdifferentials

subgradient

$$\partial f(x) = \{ g \in \mathbb{R}^d : \forall y, \quad f(y) \geq f(x) + \langle g, y - x \rangle \}$$

$\neq \emptyset$ if f convex (and proper...) ... $= \{\nabla f(x)\}$ if f differentiable

26a. Convex Learning Problems

Monday, April 1, 2024 8:59 AM

Fermat's Rule

$x \in \operatorname{argmin} f$ iff $0 \in \partial f(x)$

generalizes solving $f'(x) = 0$

Constraints

$\min_{x \in C} f(x) \overset{\text{equiv.}}{\iff} \min f(x) + \iota_C(x)$

indicator function $= \begin{cases} 0 & x \in C \\ +\infty & x \notin C \end{cases}$

So... can still apply Fermat's Rule!

⚠ Not the usual 1-0 indicator!

Convex Optimization Problems

$\min_{x \in C} f(x)$ is a "convex problem" if 1) f is a convex function
2) C is a convex set
($\Rightarrow \iota_C$ is a convex function)

Thm: In a convex problem, all local minimizers (or stationary pts) are global!

proof via pictures and epigraph

Myth: Convex problems have unique solutions

No, consider $f(x) \equiv 0$. Need strict convexity to ensure uniqueness

Variant: It doesn't matter where you initialize from for cvx problems

Not true for 2 reasons: 1) lack of uniqueness
2) you rarely find the exact soln, you just get close

...but a grain of truth.

Myth: Convex problems are easy to solve

No, let f be nonconvex (and intractable to minimize), then lsc cvx envelope f^{**} has some minimizers so is just as intractable

Lipschitz

$\varphi: \mathbb{R}^d \to \mathbb{R}^k$ is ρ-Lipschitz (value depends on choice of norm) if $\forall x, y \quad \|\varphi(x) - \varphi(y)\| \leq \rho \|x - y\|$

(if φ' exists, this is same as requiring φ' bounded Taylor Remainder Thm)

26b. Convex Learning Problems

Monday, April 1, 2024 9:12 AM

β-smooth (aka β strongly smooth)

$f: \mathbb{R}^d \to \mathbb{R}$ is β-smooth if ∇f exists and is β-Lipschitz (by default, wrt Euclidean norm)

Thm f β-smooth then $x \mapsto f(Ax+b)$ is $\|A\|^2\beta$ smooth

μ-strongly convex

$f: \mathbb{R}^d \to \mathbb{R}$ is μ-strongly cvx (wrt Eucl. norm) if

$x \mapsto f(x) - \mu/2 \|x\|_2^2$ is convex

Thm If $f: \mathbb{R}^d \to \mathbb{R}$ has a 2nd derivative $\nabla^2 f$, then if $\forall x$

$0 \le \nabla^2 f(x) \Rightarrow f$ is convex

$\mu I \le \nabla^2 f(x) \Rightarrow f$ is μ-strongly convex

$\nabla^2 f(x) \le \beta \cdot I \Rightarrow f$ is β-smooth

$A \le B \iff B - A \succcurlyeq 0 \iff B-A$ is positive semidefinite (psd)

$\iff$ all eigenvalues of $B-A$ are non-neg.

Loewner partial order "$\le$"

Sub-optimality

If $\min f(x)$ is convex, smooth & unconstrained w/ unique solution x^*, then TFAE:

1) $\|x - x^*\| = 0$
2) $f(x) - f(x^*) = 0$
3) $\|\nabla f(x)\| = 0$ ← this one is useful for nonconvex problems

In practice this almost never happens. Instead, we get

1) $\|x - x^*\| \le \varepsilon_1$
2) $f(x) - f(x^*) \le \varepsilon_2$
3) $\|\nabla f(x)\| \le \varepsilon_3$

under conditions on f, we can relate the ε's to each other

see handout

See my notebook for proofs

eg. If f β-smooth, $\|\nabla f(x)\|^2 \le 2\beta (f(x) - f(x^*))$

"self-bounded" [SS] if $f(x) \ge 0 \ \forall x$, so $\|\nabla f(x)\|^2 \le 2\beta f(x)$

If f μ-strongly cvx, $f(x) - f(x^*) \le \frac{1}{2\mu}\|\nabla f(x)\|^2$ and $\|x - x^*\|^2 \le \frac{2}{\mu}(f(x) - f(x^*))$

26c. Convex Learning Problems

Wednesday, April 3, 2024 9:10 AM

Optimization "meta-rules"

1) $\min_x f(x) = -\max_x -f(x)$ and define f is concave if $-f$ is convex

$\operatorname{argmin}_x f(x) = \operatorname{argmax}_x -f(x)$

2) if φ is monotonic increasing on range(f) then

$\operatorname{argmin} f(x) = \operatorname{argmin} \varphi(f(x))$

ex: $\min \|Ax-b\|$ and $\min \frac{1}{2}\|Ax-b\|^2$ are basically the same

3) $\min_x \min_y f(x,y) = \min_y \min_x f(x,y) = \min_{x,y} f(x,y)$

3') $\max_y \min_x f(x,y) \le \min_x \max_y f(x,y)$ "weak duality"

but usually not "="

4) $\min_x f(x)+g(x) \neq \left(\min_x f(x)\right) + \left(\min_x g(x)\right)$

4') $\min_x f(x)+g(x) \geq \left(\min_x f(x)\right) + \left(\min_x g(x)\right)$

5) $\min_{x \in C} f(x) \le \min_{x \in D} f(x)$ if $D \subseteq C$ "relaxation"

26d. Convex Learning Problems

Wednesday, April 3, 2024 9:20 AM

Connections to Learning

$Z = X \times Y$ (features, labels)

We'll choose $h \in \mathcal{H}$, and pick a loss $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$

Assume $\mathcal{H}$ can be parameterized and is isomorphic to a subset of $\mathbb{R}^d$, with $w \in \mathbb{R}^d$ and (re-using notation) $\mathcal{H} \subseteq \mathbb{R}^d$

We say a learning problem $(\mathcal{H}, Z, \ell)$ is convex if

1) $\mathcal{H} \subseteq \mathbb{R}^d$ is a convex set (hopefully also closed)

2) $\forall z \in Z$, $w \mapsto \ell(w, z)$ is a convex function

Lemma 12.11

If $(\mathcal{H}, Z, \ell)$ is a convex learning problem then its ERM problem is a convex optimization problem

proof: $\hat{L}_S(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, z_i)$ is convex, $w \in \mathcal{H}$ is a convex constraint.

Not all convex learning problems are PAC learnable

Ex. (Ex 12.8, 12.9 [SS]) 1D linear regression over $\mathcal{H} = \mathbb{R}$ (or $\mathcal{H} = [-1, 1]$) isn't PAC learnable

BACKGROUND: "EVT" A cts fcn over a compact set achieves its sup/inf.

Need more assumptions, e.g.

(1) "Cvx-Lipschitz-bdd" $\mathcal{H}$ is convex and bounded (eg. $\mathcal{H} = \{w : \|w\|_2 \le B\}$)

and $\forall z \in Z$, $w \mapsto \ell(w, z)$ is convex and ρ-Lipschitz

Note: hence loss is bounded (though possibly not uniformly in z)

$$|\ell(w, z)| \le |\ell(0, z)| + \underbrace{|\ell(w, z) - \ell(0, z)|}_{\le \rho \cdot \|w - 0\| \le \rho \cdot B}$$

(2) "Cvx-Smooth-bdd" $\mathcal{H}$ is convex and bounded (as in (1))

and $\forall z \in Z$, $w \mapsto \ell(w, z)$ is convex and β-smooth and non-negative

(if $\mathcal{H}$ is closed, then this is a stronger assumption than (1) since

$\nabla \ell$ Lipschitz $\Rightarrow \nabla \ell$ cts $\Rightarrow^{\mathcal{H}}$ bounded over $\mathcal{H}$ via compactness $\Rightarrow \ell$ is Lipschitz)

26e. Convex Learning Problems

Wednesday, April 3, 2024 9:48 AM

Ex:

$\mathcal{H} = \{ w \in \mathbb{R}^d : \|w\|_2 \le B \}$

$\mathcal{X} = \{ x \in \mathbb{R}^d : \|x\|_2 \le \rho \}$, $\mathcal{Y} = \mathbb{R}$

then

- $\ell(w, (x,y)) := |\langle w, x \rangle - y|$ is cvx-Lipschitz-bdd w/ parameters B and ρ
- $\ell(w, (x,y)) := \frac{1}{2}(\langle w, x \rangle - y)^2$ is cvx-smooth-bdd w/ parameters B and ρ^2

  Since $\nabla \ell = (\langle w, x \rangle - y) \cdot x$

  $\nabla^2 \ell = xx^T$, $\|xx^T\| = \|x\|_2^2 \le \rho^2$ (↑ spectral, ↖ Eucl.)

  ↑ Book says ρ in Ex 12.11. That's wrong.

For classification, 0-1 loss isn't continuous...

... so not Lipschitz nor β-smooth

The "hack" is to find a **surrogate loss** that is nice (smooth) and upper bounds original loss (so $L_D(h) \le L_D^{\text{surrogate}}(h)$)

eg. $\ell^{0\text{-}1}(w, z) := \mathbb{1}_{y \neq \text{sign}(\langle w, x \rangle)} = \mathbb{1}_{\underbrace{y \cdot \langle w, x \rangle}_{\alpha} \le 0}$

a common surrogate is

$\ell^{\text{hinge}}(w, z) := \max(0, 1 - \overbrace{y \langle w, x \rangle}^{\alpha})$

eg.