Lecture 18: end of margins & opt

Announcements

* Any general opt questions, including "why didn't we cover ..." (at any point in lecture).

Original (convex) opt lectures, some focus on "implicit regularization"

* E.g., concretely, hw2 p3.

Last two lectures: minimum norm solutions (via margin characterization).

Last time: (linear models)

* hard margin: $\gamma(w) = \frac{\min_i y_i x_i^T w}{\|w\|} = \min_i y_i x_i^T \left(\frac{w}{\|w\|}\right)$
* "smoothed"/"soft margin": $\tilde\gamma(w) = \frac{-\ln \sum \exp - y_i x_i^T w}{\|w\|}$
* GF on $\ln\sum\exp$: $\gamma(w(t)) \geq \tilde\gamma(w(t)) \geq \max_{\|u\|\leq 1} \gamma(u) - O\left(\frac{1}{t}\right)$

This time: (L-homogeneous models)

$f(x_i; w)$ L-homogeneous: $f(x_i; cw) = c^L f(x_i; w)$, $c \geq 0$

$\langle w, v\rangle = L f(x_i; w)$, $v \in \partial f(x_i; w)$

$m_i(w) = y_i f(x_i; w)$

* hard margin: $\min_i m_i\left(\frac{w}{\|w\|}\right) = \frac{\min_i m_i(w)}{\|w\|^L}$ ($\gamma(w)$; correct scaling)

[$\max_{\|u\|\leq 1} \min_i m_i(u)$ has maximum under various conditions (e.g., cont)
$\Updownarrow$
$\max_{u \in \mathbb{R}^p} \min_i m_i\left(\frac{u}{\|u\|}\right)$]

* "smoothed"/"soft" margin: $\tilde\gamma(w) = \frac{-\ln\sum\exp(-m_i(w))}{\|w\|^L}$

* GF on $\ln\sum\exp$

we'll show $t \mapsto \tilde\gamma(w(t))$ non decreasing $\forall t \geq t_0$, $\tilde\gamma(w(t_0)) > 0$ (Kaifeng Lyu, Jian Li) [proof by Ziwei]

very nice paper

Proposition. Suppose $\exists w$ $\tilde\gamma(w) > 0$.
Then $\inf_{v\in\mathbb{R}^p} \sum \exp(-m_i(v)) = 0$
(and it is not attained).

Proof. $0 < \frac{-\ln\sum\exp(-m_i(w))}{\|w\|^L}$

$\Rightarrow \sum\exp(-m_i(w)) < 1$

$\Rightarrow \forall i:\ m_i(w) > 0$. let $\varepsilon = \min_i m_i(w) > 0$

$0 \leq \inf_{v\in\mathbb{R}^p} \sum\exp(-m_i(v)) \leq \liminf_{c>0} \sum\exp(-m_i(c\cdot w))$

$= \liminf_{c>0} \sum\exp(-c^L \cdot m_i(w))$

$\leq \liminf_{c>0} \sum\exp(-\underbrace{c^L}_{\to 0}\cdot\underbrace{\varepsilon}_{>0}) = 0.$ //

Theorem (Lyu & Li '19, simplification by Ziwei Ji)
[conditions giving chain rule]
Suppose $\dot w(t) \in \partial \ln\sum\exp(-m_i(w(t)))$ $\forall t \geq 0$ a.e.
Suppose m_i L-homogeneous, $\exists t_0$, $\tilde\gamma(w(t_0)) > 0$.
Then $\forall t \geq t_0$, $\tilde\gamma(w(t)) \geq \tilde\gamma(w(t_0))$.

Remark. Assuming all chain rules.

Proof. Note $\forall i$: $m_i(w(t_0)) > 0$.

Plan: show $\frac{d}{dt}\tilde\gamma(w(t)) \geq 0$.

define $u(t) = -\ln\sum\exp(-m_i(w(t)))$
$v(t) = \|w(t)\|^L$

Note $\frac{d}{dt}\tilde\gamma(w(t)) = \frac{\dot u(t) v(t) - u(t)\dot v(t)}{v(t)^2}$.

Crucial fact: $\langle w(t), \dot w(t)\rangle$ ($\dot w(t) = -\partial\ln\sum\exp(-m_i(w))$)

$= \left\langle w(t), \frac{\sum_i \exp(-m_i(w))\,\partial m_i(w)}{\sum\exp(-m_i(w(t)))}\right\rangle$ (chain rule)

$= \sum_i \frac{\exp(-m_i(w))}{\sum_j \exp(-m_j(w))} \langle w(t), \partial m_i(w)\rangle$

$= \sum_i \frac{\exp(-m_i(w))}{\sum_j\exp(-m_j(w))} L\, m_i(w)$, where $m_i(w) = -\ln\exp(-m_i(w)) \geq -\ln\sum_k\exp(-m_k(w)) = u(t)$

$\geq \left(\sum_i \frac{\exp(-m_i(w))}{\sum_j \exp(-m_j(w))}\right) u(t)\cdot L$

$= L\cdot u(t)$. (★)

lower bounded by what we'd get if $u(t)$ were homogeneous.

Can use this to upper & lower bound $\dot u(t)$, $\dot v(t)$
& get $\dot u(t) v(t) - u(t)\dot v(t)$ "mechanically"

loss of global optimality:

$\dot u(t) = \langle \partial -\ln\sum\exp(-m_i(w)), -\partial\ln\sum\exp(-m_i(w))\rangle$

$= \|\dot w(t)\|^2$

homogeneous case:

$\geq \|\dot w(t)\| \left\langle \frac{w(t)}{\|w(t)\|}, \dot w(t)\right\rangle$

$\overset{★}{\geq} \|\dot w(t)\| \cdot L\cdot u(t)\cdot\frac{1}{\|w(t)\|}$.

linear case:

$\|\dot w(t)\|^2 = \|\dot w(t)\|\cdot\sup_{\|u\|=1}\langle\dot w(t), u\rangle$

$\geq \|\dot w(t)\|\cdot\langle\dot w(t), u\rangle$

$\geq \|\dot w(t)\|\cdot\gamma$ (uses properties of linear hardmargin) //

Remark (omitted topics).

* Escaping saddle points (typically by adding Gaussian noise to SGD by Rong Ge & friends.)
* Mean field optimization analysis (Song Mei, Andrea Montanari, & friends)
  Personally not sure of precise relationship to NTK; flavor is much different.
* Landscape analysis ("mode connectivity" by Ge et al, "all local optima are global" Tengyu Ma & friends)

$\sigma(\sigma(\ell)) - \sigma(-\ell)$