# Lecture 17: Nonsmooth & margins.

Announcements

* HW2 tomorrow (no late homework!)

Recap: nonsmoothness:

① "Clarke differential"

$$\partial f(w) := \mathrm{conv}\left( \left\{ \lim \nabla f(w_s) : \begin{array}{l} w_s \to w \\ \nabla f(w_s) \text{ exists} \\ \lim \nabla f(w_s) \text{ exists} \end{array} \right\} \right)$$

key properties (A) exists everywhere if $f$ locally Lip
(B) has a reasonable analog to $\nabla f$.
caveat: I don't know it real.

② Positive homogeneity:

$f$ is $L$-positive-homogeneous
$\forall c \geq 0 \quad f(c \cdot x) = c^L f(x)$.

key properties: (A) reasonable abstraction of ReLU
(B) Gradient interaction:
$\forall v \in \partial f(w) : \langle v, w \rangle = L \cdot f(w)$.

Plan for this week:

(I) one more clarke consequence
(II) "strong" implicit regularization property of GD.
  (i) linear case
  (ii) multilayer $L$-homogeneous (with Clarke flow).
  ("beyond NTK").
(III) Optimization topics I neglected.

**Proposition** ("norm preservation", Du-Hu-Lee, '18; Proof "suggested" Ziwei.)

Suppose $w(t)$ given by Clarke flow of

$$\hat{R}(w) = \frac{1}{n}\sum_k \ell(y_k f(x_k; w)),$$

where $f$ is 1-homogeneous in each layer, and suppose the following chain rule:

$$\forall u \in \partial \hat{R}(w) \Rightarrow u = \frac{1}{n}\sum_k \underbrace{\ell'(y_k f(x_k; w)) y_k}_{\ell_k'(w)} u_k, \quad u_k \in \partial f(x_k; w)$$

Then $\forall t \geq 0$, $\forall i, j$,

$$\frac{1}{2}\|W_i(t)\|^2 - \frac{1}{2}\|W_i(0)\|^2 = \frac{1}{2}\|W_j(t)\|^2 - \frac{1}{2}\|W_j(0)\|^2.$$

Proof:

$$\frac{1}{2}\|W_i(t)\|^2 - \frac{1}{2}\|W_i(0)\|^2$$

$$= \int_0^t \frac{d}{ds}\frac{1}{2}\|W_i(s)\|^2 \, ds = \int_0^t \langle W_i(s), \dot{W}_i(s)\rangle \, ds$$

$$= \int_0^t \left\langle W_i(s), \frac{1}{n}\sum_k \ell_k'(w(s)) u_{k,i}(s)\right\rangle ds$$

$$= \int_0^t \frac{1}{n}\sum_k \ell_k'(w(s)) \langle W_i(s), u_{k,i}(s)\rangle \, ds$$

$$= \int_0^t \frac{1}{n}\sum_k \ell_k'(w(s)) f(x_k; w(s)) \, ds,$$

which is independent of $i$.

## "Strong" implicit regularization.

Now we'll show GD → minimum norm solutions (not just small norm).

Our setting

* exponential or logistic loss
  ↳ $\ln(1+\exp(-z))$
* GF/CF.

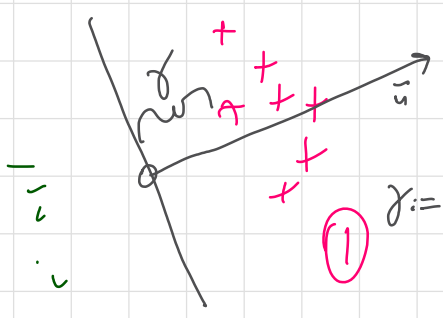
Other settings

* Squared loss, GF/GD.
  (Some authors: Gunasekar, Srebro; Cohen, Arora; )

---

Today's plan: linear case.

Here, minimum norm solution = "maximum margin solution".




① $\gamma := \max_{\|u\|\le 1} \min_i y_i x_i^T u$.

margin

max margin

② "SVM form": $\min \frac{1}{2}\|v\|^2$ s.t. $y_i x_i^T v \ge 1 \ \forall i$

$u$ is ① $\iff$ $\frac{u}{\gamma}$ is optimal for ②

Connection to exploss:

$\min: y_i x_i^T u = -\max_i -y_i x_i^T u \approx -\ln \underbrace{\sum \exp(-y_i x_i^T w)}_{\text{empirical risk}}$

In fact:

$$\gamma = \max_{\|u\|\le 1} \min_i y_i x_i^T u = \max_{r\ge 0} \max_{\|u\|\le 1} \min_i -r \ln \sum \exp\left(\frac{-y_i x_i^T u}{r}\right).$$

**Theorem** (Telgarsky '13; Gunasekar et al. '18).

Let $w(t)$ be given by $\dot{w}(s) = \nabla_w \ln \sum \exp(-z_i^T w)$, with $w(0)=0$. Then ($L_{x,y}$)

$$\gamma \ge \frac{\min z_i^T w(t)}{\|w(t)\|} \overset{①}{\ge} \frac{-\ln \sum \exp(-z_i^T w(t))}{\|w(t)\|} \ge \tilde{\gamma}(t) \ge \gamma - \frac{\ln n}{t\gamma - (\ln n)/\gamma}.$$

Remark.

Note $\nabla_v \ln \sum \exp(v) = \frac{\nabla_v \sum \exp(v)}{\sum \exp(v)}$ ?

i.e., both point in same direction.

$\Rightarrow$ Can show GF on both gives same path. ("time rescaling")

Proof. $\min_i z_i^T w(t) = -\max_i -z_i^T w(t) \ge -\ln \sum \exp(-z_i^T w(t)) \le u(t)$

$$\tilde{\gamma}(t) := \frac{u(t) - u(0) + u(0)}{\|w(t)\|} = \frac{u(0)}{\|w(t)\|} + \frac{\int_0^t \dot{u}(s)\,ds}{\|w(t)\|} \ge \gamma \text{ by } (\bigstar)$$

lose this step ↯ homogeneous

Note: $\|\dot{w}(s)\| \ge \langle \dot{w}(s), u \rangle$

$$= \langle -\nabla \ln \textstyle\sum \exp(-z_i^T w(s)), u \rangle$$

$$= \frac{\sum \exp(-z_i^T w(s))\, z_i^T u}{\sum \exp(-z_i^T w(s))}$$

$$\ge \gamma \frac{\sum \exp(-z_i^T w(s))}{\sum \exp(-z_i^T w(s))} = \gamma.$$

$$\Rightarrow \int_0^t \dot{u}(s)\,ds = \int_0^t \langle \nabla_w \ln \textstyle\sum \exp(-), \dot{w}(s) \rangle$$

$$= \int_0^t \|\dot{w}(s)\|^2 ds \ge \gamma \int_0^t \|\dot{w}(s)\|\, ds$$

$$\ge \gamma \left\| \int_0^t \dot{w}(s)\,ds \right\| = \gamma \|w(t) - w(0)\|$$

$$= \gamma \|w(t)\|.$$

$$\|w(t)\|\gamma \ge \|w(t)\|\,\tilde{\gamma}(t) = \underbrace{-\ln \textstyle\sum \exp(-z_i^T w(s))}_{u(t)}$$

$$= u(0) + \int_0^t \dot{u}(s)\,ds = u(0) + \int_0^t \|\dot{w}(s)\|^2 ds$$

$$\ge \underbrace{u(0)}_{-\ln n} + t\gamma^2 \quad \blacksquare$$