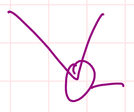
Lecture 19: GF descent guarantees; positive homogeneity

Recall:

Clarke differential $\partial f(w) = \operatorname{conv}\left\{ \lim_{i\to\infty} \nabla f(w_i) : \begin{array}{l} w_i \to w \\ \nabla f(w_i) \text{ exists} \\ \lim_i \nabla f(w_i) \end{array} \right\}$

E.g.  $\partial f(0) = [-1, +1]$.

key properties:
① always nonempty if f locally Lipschitz
(& agrees with subdifferential if f convex)

② flow property: $\dot{w}_t \in -\partial f(w_t)$,
$\Rightarrow$ for almost every $t \geq 0$, $\dot{w}_t = -\boxed{\bar{\partial}} f(w_t)$ — minimum norm element

and $\frac{d}{dt} f(w_t) = \underbrace{-\|\bar{\partial} f(w_t)\|_2^2}$.

③ Pytorch does not implement the clarke differential

Descent guarantees

Consider $\dot{w}_t \in -\partial \hat{R}(w_t)$. [Recall $\hat{R} = \mathcal{L} \circ H$.]

Proposition (stationary point). $\inf_{s<t} \|\bar{\partial} \hat{R}(w_s)\|^2 \leq \frac{1}{t}\left(\hat{R}(w_0) - \hat{R}(w_t)\right)$.

Remark. GD on β-smooth $\hat{R}$, $\eta \leq \frac{1}{\beta}$ $\Rightarrow$ $\min_{s<t} \|\nabla \hat{R}(w_s)\|^2 \leq \frac{1}{2t\eta}\left(\hat{R}(w_0) - \hat{R}(w_t)\right)$.

Note $t\eta$ is morally equivalent to t

Proof.
$$\hat{R}(w_t) - \hat{R}(w_0) \overset{\text{FTC}}{=} \int_0^t \frac{d}{ds} \hat{R}(w_s)\, ds \overset{\text{a.e.}}{=} \int -\|\bar{\partial} \hat{R}(w_s)\|^2 ds$$
$$\leq \int_0^t - \inf_{j<t} \|\bar{\partial} \hat{R}(w_j)\|^2 ds$$
$$= -t \cdot \inf_{j<t} \|\bar{\partial} \hat{R}(w_j)\|^2 \quad //$$

Rem. $\hat{R} = \mathcal{L} \circ H \Rightarrow \|\bar{\partial} \hat{R}(w_s)\|^2 \geq \sigma_{\min}(\partial H(w_s))^2 \cdot \|\partial \mathcal{L}(H(w_s))\|_2^2$.
$\Rightarrow$ if $\sigma_{\min}(\partial H(w_s))$ bounded below (by positive constant) then pass near a stationary point of $\mathcal{L}$.

Rem.

Prop. Consider $w \mapsto \max\{|w_1| - |w_2|, -100\}$.

Prop. If $\hat{R}$ is λ-sc
$$\Rightarrow \hat{R}(w_t) - \inf_v \hat{R}(v) \leq \left(\hat{R}(w_0) - \inf_v \hat{R}(v)\right) \exp(-2t\lambda).$$

Remark. For β-smooth GD, we had $\exp(-t\eta\lambda)$, $\eta \leq \frac{1}{\beta}$.

Proof. For almost all $s \geq 0$
$$\frac{d}{ds}\left[\hat{R}(w_s) - \inf_v \hat{R}(v)\right] = -\|\bar{\partial} \hat{R}(w_s)\|^2$$
$$\leq -2\lambda\left(\hat{R}(w_s) - \inf_v \hat{R}(v)\right),$$

[Recall Grönwall's inequality $\dot{u}_t \leq u_t \beta_t \Rightarrow u_t \leq u_0 \exp\left(\int_0^t \beta_s\, ds\right)$.]

therefore, by almost everywhere version Grönwall's inequality $\hat{R}(w_t) - \inf_v \hat{R}(v) \leq \left(\hat{R}(w_0) - \inf_v \hat{R}(v)\right)\exp(-2t\lambda)$ //

Similar curious theorem:

Theorem. Consider $\dot{w}_t \in -\partial \hat{R}(w_t)$, $\dot{v}_t \in -\partial \hat{R}(v_t)$, $\hat{R}$ is λ-sc
and w_0 & v_0 arbitrary (i.e., $w_0 \neq v_0$ is possible).
Then $\frac{1}{2}\|w_t - v_t\|_2^2 \leq \frac{1}{2}\|w_0 - v_0\|_2^2 \exp(-2t\lambda)$.

Proof. (similar) //

Remark. $\hat{R} = \mathcal{L} \circ H$, $\|\bar{\partial} \hat{R}(w_s)\|^2 \geq \sigma_{\min}(\partial H(w_s))^2 \cdot \|\bar{\partial} \mathcal{L}(H(w_s))\|$
$$\geq \sigma_{\min}(\partial H(w_s))^2\, 2\lambda\left(\hat{R}(w_s) - \inf_v \hat{R}(v)\right).$$

Proposition. (also in last lecture) Given reference point $z \in \mathbb{R}^d$, and convex $\hat{R}$,
$$\frac{1}{2t}\|w_t - z\|^2 + \hat{R}(w_t) \leq \frac{1}{2t}\|w_0 - z\|^2 + \hat{R}(z).$$

Rem. For GD, this guarantee had a constraint set.

Pf.
$$\frac{1}{2}\|w_t - z\|^2 - \frac{1}{2}\|w_0 - z\|^2 = \int_0^t \frac{d}{ds} \frac{1}{2}\|w_s - z\|^2 ds$$
$$= \int_0^t \langle \bar{\partial} \hat{R}(w_s), z - w_s \rangle ds$$
$$\leq \int_0^t \left[\hat{R}(z) - \hat{R}(w_s)\right] ds$$
$$\leq \int_0^t \left[\hat{R}(z) - \hat{R}(w_t)\right] ds. \quad //$$

Rem. $\hat{R} = \hat{\mathcal{L}} \circ H$, then
$$\langle \bar{\partial} \hat{R}(w_s), z - w_s \rangle = \langle \bar{\partial} \mathcal{L}(H(w_s)), \partial H(w_s)^T (z - w_s) \rangle$$
$$= \overbrace{\langle \partial \mathcal{L}(H(w_s)), H(z) - H(w_s) \rangle}^{\text{controlled via convexity of loss}}$$
$$- \underbrace{\langle \partial \mathcal{L}(H(w_s)), H(z) - H(w_s) - \partial H(w_s)^T (z - w_s) \rangle}_{\text{controlled via large width.}}$$

Positive homogeneity.

Definition. f is L-positive-homogeneous if $\forall c \geq 0$, $f(cx) = c^L f(x)$.

Remarks. ReLU $\sigma(z) = \max\{0, z\}$ is 1-homogeneous

$$\left(\overset{\forall c \geq 0}{\sigma(cz)} = \max\{0, cz\} = c \cdot \max\{0, z\} = c\,\sigma(z). \right)$$

ReLU network is L-homogeneous in parameters:

$$F(x; cw) = cW_L \sigma(cW_{L-1} \sigma(\cdots \sigma(cW_1 x) \cdots))$$

$$= c^L W_L \sigma(W_{L-1} \cdots W_1 x \cdots) = c^L F(x; W).$$

Attention layer is not positive homogeneous.

Norms are 1-homogeneous: $\|cx\| = |c| \cdot \|x\|$.

Homogeneous polynomial: all terms have same degree (common degree is L).

Relationship to gradients.

Proposition. Suppose g is L-positive homogeneous & locally Lipschitz. $\forall w$, $\forall s \in \partial g(w)$ then $\langle s, w \rangle = L \cdot g(w)$.

Examples. $\langle \partial \frac{1}{2}\|w\|^2, w \rangle = \langle w, w \rangle = \|w\|^2$.

should $= 2 \cdot \frac{1}{2}\|w\|^2 = \|w\|^2$

ReLU:
$$\sigma'(r)\,r = \begin{cases} r<0 & 0 \cdot r = 0 = \sigma(r) \\ r=0 & ? \cdot 0 = 0 = \sigma(r) \\ r>0 & 1 \cdot r = r = \sigma(r) \end{cases} \Bigg\} = \sigma(r).$$

ReLU network