Lecture 18: gradient flow & differential inclusion

GF

Announcement:

* hw2 ?? 50%

Gradient flow. (differentiable)

GD: $w' := w - \eta \nabla \hat{R}(w)$

$$\frac{w' - w}{\eta} = -\nabla \hat{R}(w)$$

$\eta \downarrow 0 \longrightarrow$ $\frac{d}{dt} w = \dot{w}_t = -\nabla \hat{R}(w_t)$

"Gradient flow" ODE; for GF, we assume solutions exist & are unique, meaning $(w_t)_{t \geq 0}$, $t \in \mathbb{R}$ is the "solution".

key byproduct

* GD on β-smooth $\hat{R}$

$$\hat{R}(w_{i+1}) \leq \hat{R}(w_i) - \eta\left(1 - \frac{\eta\beta}{2}\right)\|\nabla \hat{R}(w_i)\|^2$$

$$\Rightarrow \frac{\hat{R}(w_{i+1}) - \hat{R}(w_i)}{\eta} \leq -\left(1 - \frac{\eta\beta}{2}\right)\|\nabla \hat{R}(w_i)\|^2$$

* GF gives us this "for free"

$$\frac{d}{dt}\hat{R}(w_t) = \langle \nabla \hat{R}(w_t), \frac{d}{dt} w_t \rangle$$

$$= -\|\nabla \hat{R}(w_t)\|^2$$

needed continuity assumptions for existence & uniqueness

Why GF:

* "free" smoothness equality. (which implies $\hat{R}(w_t)$ is non-increasing).
* nice first thing to check

Why GF not nice:

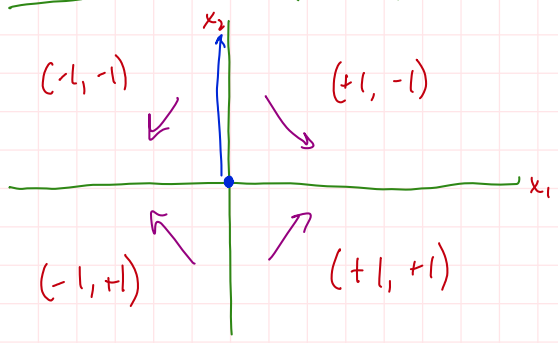
* Not an algorithm (no real "running time").
* can be hard to convert "rate" to GD (I have scary examples).
* "circuitous"

Clarke differential & GF

$\hat{R}$ diff & ($\nabla \hat{R}$ uniformly continuous)? $\Rightarrow$ $\dot{w}_t = -\nabla \hat{R}(w_t)$ (differential equation) has unique solutions

$\hat{R}$ locally Lipschitz $\Rightarrow$ $\dot{w}_t \in -\partial \hat{R}(w_t)$ (differential inclusion) may fail to have unique solutions

Ex. $x \xrightarrow{g} |x_1| - |x_2|$



Consider $\dot{z}_t \in -\partial g(z_t)$

Claim $\forall r \geq 0$

$t \in [0, r]$: $w_t = (0,0)$, $\dot{w}_t = (0,0)$

$t \in (r, \infty)$: $w_t = (0, t-r)$, $\dot{w}_t = (0,+1)$

for any a, b, $w_b - w_a$

① $r \in [a,b]$ $\Rightarrow$ $w_b - w_a = \int_a^b \dot{w}_t \, dt$

$= \int_a^r \dot{w}_s \, ds + \int_r^b \dot{w}_s \, ds$

$= 0 + \int_r^b (0,+1)$

$= 0 + (b-r)(0,+1)$.

② $a \geq r$ ③ $b \leq r$ } similar

Remarks ① (essentially univariate: consider $z \mapsto -|z|$.

② Instability also exists w.r.t. GD.

③ Kaifeng Lyu & friends "simplicity bias: ...." last appendix

"Fix": "Clarke differential inclusion is well-behaved for almost all time".

Lemma (Bolte, Lewis, J., others). Consider $\dot{z}_t \in -\partial f(z_t)$.

Suppose f is locally Lipschitz and o-minimal definable. Then ("always" satisfied for finite-width networks)

for almost every $t \geq 0$:

① chain rule: $\frac{d}{dt} f(z_t) = \langle s, \dot{z}_t \rangle$ $\forall s \in \partial f(z_t)$

② minimum norm selection: $\dot{z}_t = -\bar{\partial} f(z_t) = -\operatorname{argmin}\{\|s\| : s \in \partial f(z_t)\}$

③ $\frac{d}{dt} f(z_t) = -\|\bar{\partial} f(z_t)\|^2$.

captures main benefit of GF.

Consider again $g(x) = |x_1| - |x_2|$. Then letting x_t be a solution to Clarke flow,

$g(x_t) - g(x_0) = \int_0^t \frac{d}{ds} g(x_s) \, ds \overset{\text{(using Lemma)}}{=} -\int_0^t \|\bar{\partial} g(x_s)\|_2^2 \, ds$

($g(x_t) - g(x_0)$: assume $t \geq r$)

$= -\int_{[0,r]} \|\bar{\partial} g(x_s)\|^2 \, ds - \int_{(r,\infty)} \|\bar{\partial} g(x)\|^2 \, ds$

$= -(t - r)$.

Descent proofs.

Proposition. Suppose $\hat{R}$ is convex and $w_t \in -\partial \hat{R}(w_t)$, & $z \in \mathbb{R}^d$ arbitrary. Then $\forall t$

$$\frac{1}{2}\|w_t - z\|^2 + t\hat{R}(w_t) \le \frac{1}{2}\|w_0 - z\|^2 + t\hat{R}(z).$$

Proof.

$$\frac{1}{2}\|w_t - z\|^2 - \frac{1}{2}\|w_0 - z\|^2 = \int_0^t \frac{d}{ds}\frac{1}{2}\|w_s - z\|^2 \, ds$$

$$= \int_0^t \langle \dot{w}_s, w_s - z \rangle \, ds$$

$$= \int_0^t -\langle \bar{\partial}\hat{R}(w_s), w_s - z \rangle \, ds$$

$$\le \int_0^t \left(\hat{R}(z) - \hat{R}(w_s)\right).$$

$$\Rightarrow \frac{1}{2}\|w_t - z\|^2 + \int_0^t \hat{R}(w_s)\,ds \le \frac{1}{2}\|w_0 - z\|^2 + \int_0^t \hat{R}(z),$$ & use $\hat{R}(w_t) \le \hat{R}(w_s)$ $\forall s \le t$. //