Lecture 9

Last time
- Gauss-Newton
- Levenberg-Morrison-Marquardt

Today
- Strict saddles
- Measure zero sets
- Escaping saddles

Strict saddles

Lately we have focused on local convergence guarantees, but in practice we don't get the luxury of initializing close to a solution. This points to a natural q:

When do first-order methods converge to local/global minimizers only?

Today, we will focus on GD. We know (Lecture 2) that we can find points $\{w_k\}$ such that $\|\nabla f(w_k)\| \to 0$.

The guarantee we will prove is based on the idea that we can avoid

the following saddles.

Def: We say that $w_\#$ is a strict saddle point if

$$\nabla f(w_\#) = 0 \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(w_\#)) < 0.$$

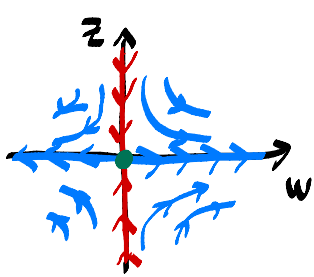
The intuitive idea is that the least eigenvector gives a direction that pushes away from $w_\#$.

Example: Consider the 2D function

$$f(w, z) = -\frac{\alpha}{2} w^2 + \frac{\beta}{2} z^2.$$

GD takes the form

$$w_{k+1} = (1 + \eta\alpha) w_k \quad \text{and} \quad z_{k+1} = (1 - \eta\beta) z_k.$$



If we follow the flow $\dot{x} = -\nabla f(x)$, only a "small" set gets trapped at zero.

Measure zero sets

To understand what "small" means, we need two measure-theoretic ideas.

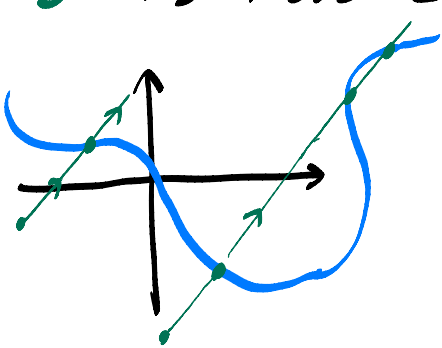
Def: Let $S \subseteq \mathbb{R}^d$ be a subset, we say that S has measure zero if for all $\varepsilon > 0$ there exists a countable sequence of balls $B_1, B_2, \ldots$ with $S \subseteq \bigcup_{i=1}^{\infty} B_i$ such that

$$\sum_{i=1}^{\infty} \text{vol}(B_i) \leq \varepsilon$$

⊣

Theorem (Fubini): Let $S \subseteq \mathbb{R}^d$ closed and $v \in \mathbb{R}^d \setminus \{0\}$. Suppose that for all $x \in \mathbb{R}^d$ we have that the intersection $(x + \text{span}\{v\}) \cap S$ is countable. Then, S has measure zero. ⊣

Finite is enough for us.



We can prove that proper subspaces and surfaces have measure zero.

Escaping strict saddles

We are now ready to state our core result for today.

Theorem: Let $f: \mathbb{R}^d \to \mathbb{R}$ be a C^2 function with an L-Lipschitz gradient. Suppose we ran (GD) with $\eta \in (0, 1/L)$ and w_0 is picked at random wrt to a distribution with a density. Then, for any strict saddle $w_\#$ we have

$$\mathbb{P}\left(\lim_{k \to \infty} w_k = w_\#\right) = 0.$$

Remarks:

1. We don't assume (GD) converges.
2. If the strict saddles form a countable set, we can avoid all of them with probability one.
3. If, additionally, all critical points

are either minimizers or strict saddles, then, if the iterates converge, they go to minimizers.

4. This is only asymptotic. In the worst case it might take $\Omega(\exp(d))$ number of iterations (Du et al. 2017).

5. Non asymptotic guarantees require adding stochastic noise to the iterates (Jin et al. 2017).

Today we will prove this result assuming some lemmas that we will prove next class.

Proof:

WLOG, suppose $w_\# = 0$ and

$$\nabla^2 f(0) = \begin{pmatrix} D_- & 0 \\ 0 & D_+ \end{pmatrix}.$$

D_-: $m \times 0$, D_+: $d - m$

where $D_- \prec 0$ and $D_+ \succeq 0$. Define the GD map as $T(x) = x - \eta \nabla f(x)$. Then

$$DT(0) = I - \eta \nabla^2 f(0) = \begin{pmatrix} I - \eta D_- & \\ & I - \eta D_+ \end{pmatrix}.$$

Define

$$\gamma = \lambda_{\min}(-D_-) > 0.$$

By continuity of $\nabla^2 f(\cdot)$ there exists $r > 0$ such that

$$\|\nabla^2 f(w) - \nabla^2 f(0)\|_{op} \le \frac{\gamma}{4} \quad \forall w \in \underbrace{\overline{B}_r(0)}_{K}$$

$$\Updownarrow$$

$$\|DT(w) - DT(0)\|_{op} \le \frac{\eta\gamma}{4} \quad \forall w \in \overline{B}_r(0).$$

Finally set

$$S = \{ w \in K \mid T^k(w) \in K, \forall k \}$$

$$= \bigcap_{k=0}^{\infty} (T^k)^{-1}(K).$$

This set contains converging trajectories and more. Let

$$\mathcal{C}(w_*) = \{ w_0 \mid T^k(w_0) \to w_* \}.$$

We want to show that $\mathcal{C}(w_\#)$ has measure zero. Any trajectory that converges to $w_\#$ eventually stays in S, so

$$(\star) \qquad \mathcal{C}(w_\#) \subseteq \bigcup_{j=0}^{\infty} (T^j)^{-1}(S).$$

We use the following results.

Lemma (Null): The set S has zero measure. ⊣

(Next class)

Lemma (Inverse Lipschitz): For any pair $w, y \in \mathbb{R}^d$, we have

$$\|T(w) - T(y)\| \geq (1 - \eta L)\|w - y\|.$$

Hence, T is injective and its inverse T^{-1} is $(1-\eta L)^{-1}$-Lipschitz.

Proof: Expanding

$$\|T(w) - T(y)\| = \|w - y - \eta(\nabla f(w) - \nabla f(y))\|$$

Reverse triangle inequality →
$$\geq \|w-y\| - \eta\|\nabla f(w) - \nabla f(y)\|$$

f is L-smooth →
$$\geq (1 - \eta L)\|w - y\|. \qquad \square$$

Next class

Lemma (-1) (Image of a null): Suppose $F: \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz, and $\mathcal{A}$ has zero measure. Then $F(\mathcal{A})$ has zero measure.

By Lemma (Null), S has zero measure. By Lemmas (L) and (-1), $(T^j)^{-1}(S)$ has zero measure for all j. Then, their countable union has zero measure. To see why, take $\varepsilon > 0$, and cover the jth set by balls $\{B_i^j\}_{i=1}^{\infty}$ of volume less than $\varepsilon/2^{j+1}$, then

(♡) $\bigcup_{j=1}^{\infty} (T^{-1})^j(S) \subseteq \underbrace{\bigcup_{j=1}^{\infty} \bigcup_{i=1}^{\infty} B_i^j}_{\mathcal{C}_\varepsilon}$ and $\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \mathrm{vol}(B_i^j) \leq \sum_{j=0}^{\infty} \frac{\varepsilon}{2^{j+1}} = \varepsilon.$

Thus, (★) and (♡) show that $\mathcal{C}(w_\#)$ has measure zero. Further, for any $\varepsilon > 0$, there exists an $M > 0$ s.t.

(·) $\int_{\|p\| \geq M} p(w)\, dw \leq \varepsilon/2.$ (why?).

Then,

$$\mathbb{P}\left(\lim_{k\to\infty} w_k = w_\#\right)$$

$$= \int_{\mathcal{C}(w_\#)} p(w)\, dw$$

$$\leq \int_{\mathcal{C}(w_\#)} \mathbb{1}_{\{p<M\}}(w)\, p(w)\, dw + \int_{\mathbb{R}^d} \mathbb{1}_{\{p\geq M\}}\, p(w)\, dw$$

$$\overset{(*)}{\leq} M \int_{C_{\varepsilon/2M}} dw + \varepsilon/2$$

(*) ↘ Measure zero ↗

$$\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since ε was arbitrary, the conclusion follows. □