Constrained optimization

we will focus on constrained opt. problems today:

$$(P) \qquad \min_{x \in \mathbb{R}^d} f(x) \quad \text{s.t.} \quad g_i(x) \le 0 \quad \forall i = 1, \ldots, n.$$

We will assume each of f, g_i are C^1 (continuously differentiable)
Similar results hold if we assume they are locally Lipschitz, but requires more technical arguments w/ Clarke subdifferentials etc. (see Lyu-Li '19)

Def A point $x \in \mathbb{R}^d$ is called feasible for (P) if $g_i(x) \le 0$ $\forall i = 1, \ldots, n$.
A feasible point x is called a KKT point (Karush-Kuhn-Tucker) if x satisfies the KKT conditions: $\exists \lambda_1, \ldots, \lambda_n \ge 0$ s.t.

1. $\nabla f(x) + \sum_{i=1}^{n} \lambda_n \nabla g_i(x) = 0.$
2. $\forall i = 1, \ldots, n,$ $\lambda_i(x) \cdot g_i(x) = 0.$

- Global minimum of (P) may not be a KKT point.
- Under certain "regularity conditions", we can guarantee this.

Def. (Mangasarian-Fromovitz constraint qualification [MFCQ])

For feasible point x of (P), (P) satisfies MFCQ @ x if there exists $v \in \mathbb{R}^d$ s.t. for all $i \in [n]$ s.t. $g_i(x) = 0$,

$$\langle \nabla g_i(x), v \rangle > 0.$$

Theorem If a feasible point $x \in \mathbb{R}^d$ of (P) satisfies MFCQ, and if x is a local minimum of (P), then x satisfies the KKT conditions for (P).

Proof is somewhat involved. Reference: Andreasson, Evgrafov, Patriksson Ch. 5.

Example. Consider $f(x;\theta)$ s.t.:

① $f(x,\theta)$ is C^1 fcn of θ for every $x \in \mathbb{R}^d$,

② $f(x;\theta)$ is L-positively homogeneous for some $L > 0$: $f(x;\alpha\theta) = \alpha^L f(x;\theta)$ $\alpha \geq 0$.

Consider (P) $\min \|\theta\|_2^2$ s.t. $y_i f(x_i;\theta) \geq 1$ $\forall i = 1, \dots, n$.

Corresponds to constraints $g_i(\theta) := 1 - y_i f(x_i;\theta) \leq 0$.

Then every feasible point satisfies MFCQ:

- let θ be s.t. $g_i(\theta) = 0 \iff 1 = y_i f(x_i; \theta)$.
- want some v st $\langle v, \nabla g_i(\theta) \rangle > 0$.

$\nabla g_i(\theta) = -y_i \nabla f(x_i; \theta)$. By homogeneity, taking $v := -\theta$ yields

$$\langle v, \nabla g_i(\theta) \rangle = \langle \theta, y_i \nabla f(x_i; \theta) \rangle = y_i \cdot L f(x_i; \theta) = L > 0.$$

Putting this together:

<u>Proposition</u> Let $f(x; \theta)$ be L-homogeneous and C^1 in θ for every x. Then every local minimum of

(P) $\quad \min \|\theta\|_2^2 \quad$ s.t. $\quad y_i f(x_i; \theta) \geq 1 \quad$ for $i = 1, \ldots, n$,

is a KKT point of problem (P).

Many examples of neural nets satisfy this.

- Linear classifiers: $f(x; \theta) = \langle \theta, x \rangle$ clearly C^1, 1-homog.

- Depth-D neural nets with activations $\varphi(t) = \max(0,t)^q$, $q > 1$:
(and no bias terms)

eg 2-layer nets, $\sum_{j=1}^{m} \alpha a_j \varphi(\langle \alpha w_j, x \rangle) = \alpha \cdot \sum_{j=1}^{m} a_j \cdot \alpha^q \varphi\langle w_j, x \rangle$

$= \alpha^{q+1} \sum_{j=1}^{m} a_j \varphi(\langle w_j, x \rangle)$

Training both layers results in $(q+1)$-homog. nets.

Need $q > 1$ since $\varphi'(t) = (q-1)\max(0,t)$ if $q > 1$, but if $q = 1$ then φ' is not continuous.

- Holds for more general φ if (1) homogeneous, (2) C^1.

- Similar arg. shows 2-layer nets w/ bias terms are homog. if φ is homog., but bias terms break homogeneity for depth > 2.

We will see that gradient descent/flow on the logistic & exponential losses has an <u>implicit bias</u> towards solutions which satisfy the KKT conditions for margin maximization.

Namely:

<u>Thm</u>. [Lyu-Li'19; Ji-Telgarsky'20] Suppose ℓ is the logistic or exponential loss, $\{(x_i, y_i)\}_{i=1}^n$ training data, $y_i \in \{\pm 1\}$. Let $f(x;\theta)$ be L-homog. in θ and suppose $f(x,\theta)$ is C^2 on $\mathbb{R}^d$. Then for step size α sufficiently small, if $\exists T > 0$ st $\hat{L}(\theta^{(T)}) = \frac{1}{n}\sum_1^n \ell(y_i f(x_i;\theta^{(T)})) < \frac{1}{n}$, then

① $\hat{L}(\theta^{(t)}) \to 0$ as $t \to \infty$

② $\lim_{t\to\infty} \frac{\theta^{(t)}}{\|\theta^{(t)}\|} = \theta^*$ exists, and

$\exists \beta > 0$ s.t. $\beta \cdot \theta^*$ satisfies the KKT conditions for

$$(P) \quad \min \|\theta\|_2^2 \; : \; y_i f(x_i;\theta) \geq 1, \; \forall i = 1,..,n.$$

Thus, KKT conditions for (ℓ_2)-margin maximization characterize the limiting behavior of a large class of neural nets.

<u>Example</u> Let's write out the KKT conditions for linear classifier:

(P) $\min \frac{1}{2}\|w\|^2$ st. $y_i \langle w, x_i \rangle \geq 1$, $i = 1, \dots, n$.

for some $\lambda_i \geq 0$,

① $\nabla \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \lambda_i \nabla (1 - y_i \langle w, x_i \rangle) = 0$

$\Leftrightarrow \quad w + \sum_{i=1}^{n} \lambda_i \cdot (-y_i x_i) = 0.$

$\Leftrightarrow \quad w = \sum_{i=1}^{n} \lambda_i y_i x_i. \quad (*)$

② $\lambda_i \cdot (1 - y_i \langle w, x_i \rangle) = 0$ for all i.

Since $y_i \langle w, x_i \rangle \geq 1$, this means for every example, either:

(i) $\lambda_i = 0$ and $y_i \langle w, x_i \rangle > 1$; then (x_i, y_i) doesn't contribute to w by $(*)$.

or

(ii) $\lambda_i > 0$ and $y_i \langle w, x_i \rangle = 1$. These are "support vectors" x_i, since they lie "on the margin.

Visually:

Green examples = support vec.
These contribute to max margin.

Blue examples do not.