Chapter 9 Nonconvex functions

- Lemma 9.1 Let $f \in C^2$, with $X \subseteq \mathbf{dom}(f)$ convex, if $\|\nabla^2 f(\mathbf{x})\| \le L, \forall \mathbf{x}$, where $\|\cdot\|$ is spectual norm, then f is L-smooth.
  - Proof similar to Lemma A of Chapter 10.
- Idea For non convex function, instead of focusing on f, we focus on convergence of $\|\nabla f(\mathbf{x}_t)\|^2$ to a critical point.
- Theorem 9.2 $f \in C^2$ is L-smooth with global minimum $\mathbf{x}^\star$, then a stepsize of $\gamma = 1/L$ gives $\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{2L}{T}\left(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\right)$, and $\lim_{t\to\infty}\|\nabla f(\mathbf{x}_t)\|^2 = 0$.
  - Proof
    - sufficient descent gives $f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$, which means $\|\nabla f(\mathbf{x}_t)\|^2 \le 2L\left(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})\right)$
    - so summation and we get the first result, also if $\lim_{t\to\infty}\|\nabla f(\mathbf{x}_t)\|^2 = g > 0$ will lead to contradiction.
- Lemma 9.3(with stepsize 1/L, it cannot overshoot.) $f \in C^2$ is L-smooth, if $\nabla f(\mathbf{x}) \ne \mathbf{0}$, then update with $\gamma = 1/L' < 1/L$ will never give a critical point $\nabla f(\mathbf{x}' = \mathbf{x} - \gamma\nabla f(\mathbf{x})) \ne 0$.
  - Proof
    - By smoothness we have L-Lipschitz $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \le L\|\mathbf{x} - \mathbf{x}'\| = \frac{L}{L'}\|\nabla f(\mathbf{x})\| < \|\nabla f(\mathbf{x})\|$
    - This means $\|\nabla f(\mathbf{x}')\| \ge \|\nabla f(\mathbf{x})\| - \|\nabla f(\mathbf{x}') - \nabla f(\mathbf{x})\| > 0$.

Trajectory analysis

- Some times we can prove GD avoids saddle points and converge to global optimal

Deep Linear Neural Networks

- Objective $\|W_\ell W_{\ell-1}\cdots W_1 X - Y\|_F^2$

Width-1 DLNN

- We want when $x = 1$ then $y = 1$, this gives an objective of $f(\mathbf{x}) := \frac{1}{2}\left(\prod_{k=1}^d x_k - 1\right)^2$
- The gradient gives $\nabla f(\mathbf{x}) = \left(\prod_k x_k - 1\right)\left(\prod_{k\ne 1} x_k, \ldots, \prod_{k\ne d} x_k\right)^\top$
  - global minimum when $\prod_k x_k = 1$
  - other critical point when at least *two* x_k is zero, thay give non-minimum of $f = 1/2$.
- We want to show that from anyhwere in $X = \{\mathbf{x} : \mathbf{x} > \mathbf{0}, \prod_k \mathbf{x}_k \le 1\}$, GD converge to global minimum. However, f is not smooth in X.
- But we can later show f smooth along trajectory, then with sufficient descent, we know f always decreasing, and the starting point, we have $f < 1/2$, then never to a saddle point.
- Even in this, we still cannot prove global minimum, since X is unbounded, GD may make $\mathbf{x}$ to infinity.
- Definition 9.4 If $\mathbf{x} > 0$ componentwise, let $c \ge 1$, $\mathbf{x}$ is called c-balanced if $x_i \le c x_j$ for all $1 \le i, j \le d$
- Lemma 9.5 If $\mathbf{x} > 0$ be c-balanced with $\prod_k x_k \le 1$, then for any stepsize $\gamma > 0, \mathbf{x}' := \mathbf{x} - \gamma\nabla f(\mathbf{x})$ satisfies (i) $\mathbf{x}' \ge \mathbf{x}$ componentwise and (ii) $\mathbf{x}'$ is also c-balanced.
  - Proof
    - Set $\Delta := -\gamma\left(\prod_k x_k - 1\right)\left(\prod_k x_k\right) \ge 0$, then gradient descent gives $x_k' = x_k + \frac{\Delta}{x_k} \ge x_k$
    - We havee $x_i \le c x_j$ and $x_j \le c x_i \Leftrightarrow 1/x_i \le c/x_j$, so $x_i' = x_i + \frac{\Delta}{x_i} \le c x_j + \frac{\Delta c}{x_j} = c x_j'$
  - If we define $c \le 1$-co-balanced as $x_i \ge c x_j$ for all $1 \le i, j \le d$
  - then when $\prod_k x_k \ge 1$, then $\mathbf{x}' < \mathbf{x}$, while still $\mathbf{x}'$ is c-co-balanced.

Smoothness along the trajectory

- We can derive smoothness from bounded Hessian
- The hessian is $\nabla^2 f(\mathbf{x})_{ij} = \begin{cases} \left(\prod_{k\ne i} x_i\right)^2, & j = i \\ 2\prod_{k\ne i} x_k \prod_{k\ne j} x_k - \prod_{k\ne i,j} x_k, & j \ne i \end{cases}$
- Lemma 9.6 If $\mathbf{x} > 0$ is c-balanced, then for any subset $I \subseteq \{1, \ldots, d\}$, $\left(\frac{1}{c}\right)^{|I|}\left(\prod_k x_k\right)^{1-|I|/d} \le \prod_{k\notin I} x_k \le c^{|I|}\left(\prod_k x_k\right)^{1-|I|/d}$
  - Proof

        - For any i, we have $x_i^d \geq (1/c)^d \prod_k x_k$ so $x_i \geq (1/c)(\prod_k x_k)^{1/d}$, similarly $x_i^d \leq c^d \prod_k x_k$ so $x_i \leq c(\prod_k x_k)^{1/d}$
        - Plug in this and we get the result.
    - If c-co-balanced, $I \subseteq \{1, \ldots, d\}$, $\left(\frac{1}{c}\right)^{|I|}(\prod_k x_k)^{1-|I|/d} \geq \prod_{k \notin I} x_k \geq c^{|I|}(\prod_k x_k)^{1-|I|/d}$

- Lemma 9.7 If $\mathbf{x} > 0$ be c-balanced with $\prod_k x_k \leq 1$, then $\|\nabla^2 f(\mathbf{x})\| \leq \|\nabla^2 f(\mathbf{x})\|_F \leq 3dc^2$

    - Proof
        - For any matrix A, $\|Ax\|^2 = \sum_i (a_i^\top x)^2 \leq \sum_i (\sum_j a_{ij})^2 (\sum_j x_j)^2 = \|A\|_F^2 \|x\|_2^2$, then $\|A\| \leq \|A\|_F$
        - To bound $\nabla^2 f$, first we bound on diagnal term $|\nabla^2 f(\mathbf{x})_{ii}| = |\left(\prod_{k \neq i} x_i\right)^2| \leq c^2$
        - then for off-diagnal term $|\nabla^2 f(\mathbf{x})_{ij}| \leq |2 \prod_{k \neq i} x_k \prod_{k \neq j} x_k| + |\prod_{k \neq i,j} x_k| \leq 3c^2$
        - sum together we get $\|\nabla^2 f\|_F^2 \leq dc^4 + 9d(d-1)c^4 \leq 9d^2c^4$, QED.
    - If c-co-balanced, we can prove $\|\nabla^2 f(\mathbf{x})\| \leq \|\nabla^2 f(\mathbf{x})\|_F \leq 3d\frac{1}{c^2}$

- Lemma 9.8 (Summary of previous) If $\mathbf{x} > 0$ be c-balanced with $\prod_k x_k \leq 1$, $L = 3dc^2$, Let $\gamma := 1/L$, then GD with this lr gives $\mathbf{x}_t$ always c-balanced, and f is L-smooth along the line segment of trajectory.

    - Proof key is that smooth function never pass critical point, so every iterate, we have $\prod_k x_k \leq 1$.
    - We can prove similar result for $\prod_k x_k \geq 1$ case with similar definition of c-co-balance (Exercise 58).

- Exercise 59 there are starting point $\mathbf{x}_0$ not critical that does not converge to global minimum.

    - When $\prod_k x_k \geq 1$ and $\Delta \leq 0$, then update $x_k' = x_k + \Delta / x_k$ will lead to zero for some large learning rate.

Convergence

- Theorem 9.9 Let $c > 1$ and $\delta > 0$ such that $\mathbf{x}_0 > 0$ is c-balanced with $\delta \leq \prod_k (\mathbf{x}_0)_k < 1$, choosing stepsize $\gamma = \frac{1}{3dc^2}$, then GD satisfies $f(\mathbf{x}_T) \leq \left(1 - \frac{\delta^2}{3c^4}\right)^T f(\mathbf{x}_0)$.
    - Proof
        - By sufficient decrease $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{6dc^2}\|\nabla f(\mathbf{x}_t)\|^2$
        - while $\|\nabla f(\mathbf{x})\|^2 = 2f(\mathbf{x}) \sum_{i=1}^d \left(\prod_{k \neq i} x_k\right)^2 \geq 2f(\mathbf{x})\frac{d}{c^2}(\prod_k x_k)^{2-2/d} \geq 2f(\mathbf{x})\frac{d}{c^2}(\prod_k x_k)^2 \geq 2f(\mathbf{x})\frac{d}{c^2}\delta^2$
        - then $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{6dc^2} 2f(\mathbf{x}_t)\frac{d}{c^2}\delta^2 = f(\mathbf{x}_t)\left(1 - \frac{\delta^2}{3c^4}\right)$
- Exercise 61 Sequence $(\mathbf{x}_T)_{T \geq 0}$ in above update converge to an optimal solution $\mathbf{x}^\star$
    - Proof
        - Since $0 < x_i \leq c(\prod_k x_k)^{1/d} \leq c$, sequence is always bounded, then it has a converging subsequence $\{\mathbf{x}_{t_k}\}$.
        - By Young's inequality $(\sum_i a_i)^2 = \sum_{ij} a_i a_j \leq \sum_{ij} \frac{1}{2}(a_i^2 + a_j^2) = n\sum_i a_i^2$, this also fits for vector case, $\|\sum_i \mathbf{a}_i\|^2 \leq n \sum_i \|\mathbf{a}_i\|_2^2$
        - Let $\mathbf{a}_t = \mathbf{x}_{t_k} - \mathbf{x}_t$, then we have $\|\mathbf{x}_{t_k} - \mathbf{x}_T\|^2 \leq (T - t_k)\sum_t \|\gamma \nabla f(x_t)\|^2 \leq C \cdot (T - t_k) \cdot (f(x_{t_k}) - f(x_T))$, $f(x_{t_k}) - f(x_T)$ converge exponentially w.r.t t_k, so this term $\|\mathbf{x}_{t_k} - \mathbf{x}_T\|^2$ converge to zero.
    - We can also prove from the fact of $x_{k,t}$ is monotone....way much easier.