Convergence theory in ML

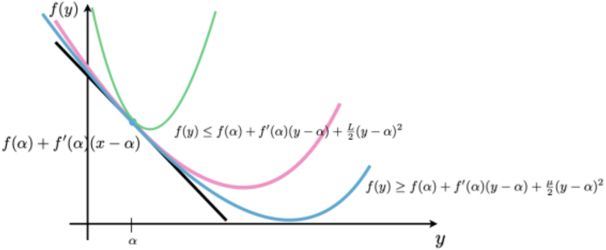
Masa Maksimovic

Rice University, Fall reading group

11^{th} September, 2023

Basic convergence results - Definitions

- $L-$smoothness - $||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||_2, \forall x, y, L > 0$
- Convexity - $f : \mathbb{R} \to \mathbb{R}$ *is an univariate convex function if,*
  $\forall \alpha \in [0, 1]$
  $f(\alpha x + (1 - \alpha y)) \leq \alpha f(x) + (1 - \alpha) f(y), \forall x, y$
- Additionally:
- Strong $(\mu-)$ convexity - *A function* $f : \mathbb{R}^p \to \mathbb{R}$ *is a strongly convex function if it is convex and, for* $\mu > 0$ *satisfies*
  $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}||x - y||_2^2, \forall x, y$

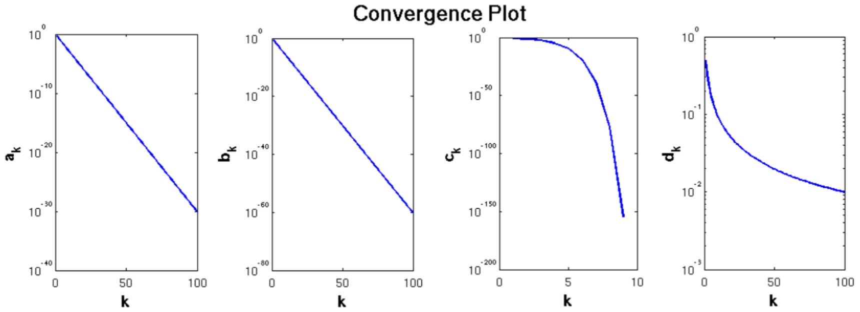


Gradient descent under L-smoothness assumption

- Idea behind gradient descent: Assuming our loss function is differentiable, we can approximate it with Taylor's expansion $f(x+\delta) = f(x) + \langle \nabla f(x), \delta \rangle + o(||\delta||_2)$
- In order to minimize f locally, we should find δ such that keeps $\langle \nabla f(x), \delta \rangle$ as small as possible (moving towards right direction).
- Therefore it is obvious that we say $\delta = -\frac{\nabla f(x)}{||\nabla f(x)||_2}$, so we get a direction with controllable step $\delta = -\eta \nabla f(x)$
- We then formally define gradient descent as following:
- *Let f be a differentiable objective with gradient $\nabla f(\cdot)$. The gradient descent method optimize f iteratively , as in* $x_{t+1} = x_t - \eta_t \nabla f(x_t), t = 0, 1...$ *where x_t is the current estimate, and η_t is the step size.*

Gradient descent under L-smoothness assumption

- Under L-smoothness assumptions we claim:
- *Assume we run gradient descent for T iterations, and we obtain T gradients, $\nabla f(x_t)$ for $t \in 0, ..., T$.* Then,
  $\min_{t \in 0,...,T} ||\nabla f(x_t)||_2 \leq \sqrt{\frac{2L}{T+1}}(f(x_0) - f(x^*))^{\frac{1}{2}} = O(\frac{1}{\sqrt{T}})$
- We have *sublinear convergence rate*.



Gradient descent under $L-$smoothness and convexity assumptions

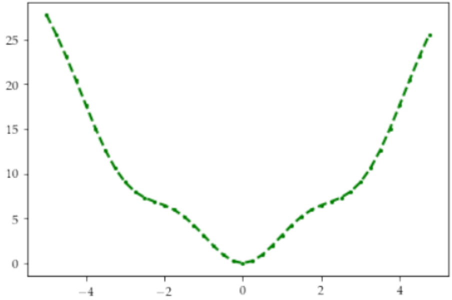
- Adding convexity we have:
- $f(x_T) - f(x^*) \leq \frac{2L(f(x_0)-f(x^*))\cdot||x_0-x^*||_2^2}{2L||x_0-x^*||_2^2+T\cdot(f(x_0)-f(x^2))} = O(\frac{1}{T})$ which is improved comparing to only $L-$smoothness assumption
- Now assuming strong $(\mu-)$ convexity we have: $||x_T - x^*|| \leq O((\frac{\kappa-1}{\kappa+1})^T) \cdot ||x_0 - x^*||_2^2$ where $\kappa := \frac{L}{\mu} > 1$, which leads us to *linear convergence rate*.
- Note: Can we achieve an even better convergence rate under $L-$smoothness and $\mu-$convexity? - Lower bounds analysis.

Gradient descent under other assumptions

- PL (*Polyak Lojasiewicz*) inequality:
- *A function f satisfies the PL inequality, if the following holds for some* $\zeta > 0$
  $\frac{1}{2}||\nabla f(x)||_2^2 \geq \zeta \cdot (f(x) - f(x^*)), \forall x$
- Assuming $L-$smoothness and PL inequality for an objective f we have:
  $f(x_T) - f(x^*) \leq (1 - \frac{\zeta}{L})^T \cdot (f(x_0) - f(x^*))$
  which leads us to *linear convergence rate* (assuming $L \geq \zeta$)

Gradient descent under other assumptions

- We have linear convergence rate without assuming convexity of an objective f!
- There is also a hiearchy of other inequalities, some of which are more restrained than others. (μ is different in each of them)



$f(x) = x^2 + 3\sin^2(x)$

When we know neither L nor μ

- In this case we need to find an adaptive step size η regardless of L and μ and theoretically justify the choice.
- There are several approaches to this problem, but we will present only *Polyak step size*
- We only assume that function f is convex (and non-smooth) and will focus on the generic (sub)gradient descent with following recursion:
  $x_{t+1} = x_t - \eta_t g_t$
- Additionally we will assume that $||g_t||_2 < G$ for some constant G.
- It appears that suitable step size is as follows:
  $\eta_t = \frac{f(x_t) - f(x^*)}{||g_t||_2^2}$
- Convergence rate:
  $\underset{t \in 0,\dots,T}{min} f(x_t) - f(x^*) \leq \frac{G||x_0 - x^*||_2}{\sqrt{t+1}} = O(\frac{1}{\sqrt{t}})$
- Note: What is the caveat?

Convergence in deep learning

- We make two assumptions: the inputs do not degenerate and the network is over-parameterized.
- Number of hidden neurons is sufficiently large - polynomial in n (number of training samples) and in L (number of layers)
- The theory applies to non-smooth ReLU activation function and to any smooth and possibly non-convex loss function.

Convergence of gradient descent

Theorem 1

For any $\epsilon \in (0,1], \delta \in (0, O(\frac{1}{L})]$. *Let* $m \geq \tilde{\Omega}((nL/\delta)^{30} * d * \log^2 \epsilon^{-1})$, $\eta = \mathcal{O}(\frac{d\delta}{n^4 L^2 m})$ *and* $\vec{\mathcal{W}}, \mathcal{A}, \mathcal{B}$ *are at random initialization. Then, with probability at least* $1 - e^{-\Omega \log m^2}$*, suppose we start at* $\vec{\mathcal{W}}^{(0)}$ *and for each* $t = 0, ..., T$,

$$\vec{\mathcal{W}}^{(t+1)} = \vec{\mathcal{W}}^{(t)} - \eta \nabla F(\vec{\mathcal{W}}^{(t)})$$

Then it satisfies

$$F(\vec{\mathcal{W}}^{(T)} \leq \epsilon) \quad \textit{for} \quad T = \mathcal{O}(\frac{n^6 L^2}{\delta^2} \log \frac{1}{\epsilon})$$

Auxiliary claims

Lemma

If $\epsilon \in (0,1]$, with probability at least $1 - nLe^{-\Omega(m\epsilon^2/L)}$ over the randomness of $\mathcal{A} \in \mathrm{R}^{m\times\sigma}$ and $\vec{\mathcal{W}} \in (\mathrm{R}^{m\times m})^L$, we have

$$\forall i \in [n], l \in \{0, 1, ...L\} : ||h_{i,l}|| \in [1-\epsilon, 1+\epsilon].$$

Theorem 2

With probability at least $1 - e^{-\Omega(m/poly(n,L,\delta^{-1}))}$ over the randomness of $\vec{\mathcal{W}}^{(0)}, \mathcal{A}, \mathcal{B}$, it satisfies for every $l \in [L]$, every $i \in [n]$, and every $\vec{\mathcal{W}}$ with $||\vec{\mathcal{W}} - \vec{\mathcal{W}}^{(0)}||_2 \le \frac{1}{poly(n,L,\delta^{-1})}$,

$$||\nabla F(\vec{\mathcal{W}})||_F^2 \le O(F(\vec{\mathcal{W}}) \times \frac{Lnm}{d}) \quad and \quad ||\nabla F(\vec{\mathcal{W}})||_F^2 \ge \Omega(F(\vec{\mathcal{W}}) \times \frac{\delta m}{dn^2}).$$

Auxiliary claims

Theorem 3

With probability at least $1 - e^{-\Omega(m/poly(n,L,\delta^{-1}))}$ *over the randomness of* $\vec{\mathcal{W}}^{(0)}$, $\mathcal{A}$, $\mathcal{B}$, *we have for every* $\vec{\mathcal{W}} \in (\mathrm{R}^{m\times m})^L$ *with* $||\vec{\mathcal{W}} - \vec{\mathcal{W}}^{(0)}||_2 \leq \frac{1}{poly(L,\log m)}$, *and for every* $\vec{\mathcal{W}}' \in (\mathrm{R}^{m\times m})^L$ *with* $||\vec{\mathcal{W}}'||_2 \leq \frac{1}{poly(L,\log m)}$:

$$F(\vec{\mathcal{W}} + \vec{\mathcal{W}}') \leq F(\vec{\mathcal{W}}) + \langle \nabla F(\vec{\mathcal{W}}), \vec{\mathcal{W}}' \rangle$$
$$+ \frac{poly(L)\sqrt{nm\log m}}{\sqrt{d}} \cdot ||\vec{\mathcal{W}}'||_2 (F(\vec{\mathcal{W}}))^{1/2} + O(\frac{nL^2 m}{d})||\vec{\mathcal{W}}'||_2^2$$

- Main techniques used in proving claims above: properties at random initialization, stability after adversarial perturbation, gradient bound, smoothness.

Conclusion

Gradient descent in over-parametrized DNN has $\epsilon-$error solution with linear convergence rate starting from random *Gaussian* initialized weights!