Special Properties of Gradient Descent with Large Learning Rates

Denis Grachev

May 29, 2023

Abstract

This article is an overview of the work [MJS23] with some additional experiments. The work focuses on theoretical proof of escaping local minima with large learning rates in optimization minimization with some special functions.

1 Main results

Usage of large learning rate is often explained by intuition of escaping local minima. In this work a class C_l of functions is constructed wich have at least two minima ($x^\dagger$ and x_*) and with a large learning rate GD with random initialization converges to x_* almost surely, but with smaller learning rate there is strictly positive probabilty of converging to $x^\dagger$

2 Theoretical Analysis

For analysis optimization minimization problem using full-batch gradient descent with random initialization was taken:

$$f_* := \min_{x \in \mathbb{R}^d} f(x).$$

f is supposed to be L-smooth over regions of the landscape so that the gradient does not change too sharply. Also we would require sharpness of some regions of f around local minima using μ-one-point-strongly-convexity (OPSC) with respect to x_* over M.

Definition 2.1 (L-smoothness)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth $\Leftrightarrow$ f is differentiable and $\exists L : \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, $\forall x, y \in \mathbb{R}^d$

Definition 2.2 (μ-one-point-strongly-convex (OPSC) with respect to x_* over M)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-one-point-strongly-convex (OPSC) with respect to x_* over M if it is differentiable and

$$\exists \mu > 0 : \langle \nabla f(x), x - x_* \rangle \geq \mu \|x - x_*\|^2, \ \forall x \in M.$$

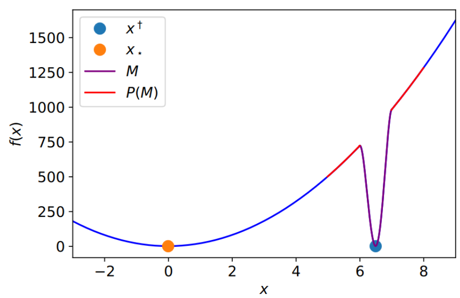
Lemma 2.1. Let f be a function that is L_{global} -smooth with a globl minimum x_*. Assume there exists a local minimum $x^\dagger$ around which

- f is $\mu^\dagger$-OPSC woth respect to $x^\dagger$ over a set M that contains $x^\dagger$ with diameter r.
- Let $P(M)$ be a ball around $x^\dagger$ with radius r_P excluding points M. f is $L < L_{\text{global}}$ -smooth in $P(M)$ and μ_*-OPSC with respect to x_*, such as $\mu^\dagger > \frac{2L^2}{\mu_*}$. r_P depends on $r, \gamma, L_{\text{global}}$.
- $\|x_* - x^\dagger\| > \tau$, where τ depends on μ_*, r, γ.

  Then using learning rate $\frac{2}{\mu^\dagger} < \gamma < \frac{\mu_*}{L^2}$ GD escape M and reach a point closer to x_* than $\|x^\dagger - x_*\| - r$ almost surely.

Proof. □

Figure 1: Illustation of regions from lemma 2.1



Theorem 2.2. Let C_l be the set of functions sudh as f is L-smooth and μ_*-OPSC with respect to the global minima x_* except n a region M that contains local minima $x^\dagger$ and satisfies lemma 2.1.

- Gradient descent initialized randomly inside M with kearning rate $\gamma < \frac{\mu^\dagger}{L^2_{\text{global}}}$ converges to $x^\dagger$ almost surely.
- Gradient descent initialized randomly in arbitary set $W : \mathcal{L}(W) > 0$ with learning rate $\frac{2}{\mu^\dagger} < gamma \leq \frac{\mu_*}{L^2}$ converges to x_* almost surely.

Proof. □

Lemma 2.3. Take gradient descent initialized randomly in set W with learning rate $\gamma \leq \frac{1}{2L}$. Let $X \subset \mathbb{R}^d$ arbitary set of points in the landscape, f is L-smooth over $\mathbb{R}^d \setminus X$. Probabilty of encountering any point of X in first T steps of gradient descent is at most $2^{(T+1)d}\frac{\mathcal{L}(X)}{\mathcal{L}(W)}$.

Proof. □

Theorem 2.4. Let X be an arbitary set of points, f is μ_*-OPSC with respect to a minima $x_* \notin X$ over $\mathbb{R}^d \setminus X$. Let $c_X := \inf\{\|x - x_*\| \mid x \in X\}$ and $r_W := \sup\{\|x - x_*\| \mid x \in W\}$. The probabilty of not encountering ant points of X during gradient descent with learning ratee $\gamma \leq \frac{\mu_*}{L^2}$ is at least $1 - \frac{r_W}{c_X}^{\frac{-d}{\log_2(1-\gamma\mu_*)}} \frac{\mathcal{L}(X)}{\mathcal{L}(W)} 2^d$ if $c_X \leq r_W$ and 1 otherwise.

Proof. □

Proposition

Take the case of SGD

$$x_{t+1} := x_t - \gamma\left(\nabla f(x_t) + \xi_t\right),$$

where ξ_t considered to be Uniform$(-\sigma, \sigma)$.

Proposition 2.5. Consider SGD on the function from figure 1, starting close to $x^\dagger$. If the learning rate is sufficiently small the iterations will never converge to x_* nor to a small region around it, regardless of the magnitude of the noise. If the learning rate is large enough and stochastic noise satisfies certain bounds, SGD will converge to x_* from any starting point.

Proof. □

References

[MJS23] Amirkeivan Mohtashami, Martin Jaggi, and Sebastian Stich. Special properties of gradient descent with large learning rates, 2023.