Special Properties of Gradient Descent with Large Learning Rates

Denis Grachev

July 25, 2023

Abstract

This article is an overview of the work [MJS23] with some additional experiments. It has been widely observed that usage of larger learning rate in stochastic gradient descent often results into better models. However theoretical reasons for this phenomena are not well understood yet. Previous studies linked it to stochastic noise in SGD. The work focuses on theoretical proof of escaping sharp minima for some special class of functions. Also it is shown that for certain starting points and loss functions GD with large learning rate has different trajectory and may lead to convergence to another minima, which is likely to be more robust.

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 which have at least two minima ($x^\dagger$ and x_*). Randomly initialized GD with large learning rate converges to x_* almost surely, but with smaller learning rate there is strictly positive probability of converging to $x^\dagger$.

2 Theoretical Analysis

For analysis 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 if 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 global minimum x_*. Assume there exists a local minimum $x^\dagger$ around which*

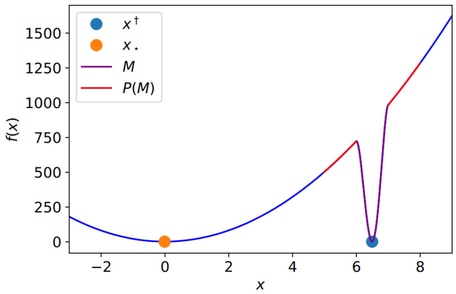
- *f is $\mu^\dagger$-OPSC woth respect to $x^\dagger$ over a set M that contains $x^\dagger$ with diameter r.*
- *Let $P(M)$ be a ball around $x^\dagger$ with radius r_P excluding points M. f is $L < L_{\text{global}}$ -smooth in $P(M)$ and μ_*-OPSC with respect to x_*, such as $\mu^\dagger > \frac{2L^2}{\mu_*}$. 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. Proof is presented in [MJS23] in appendix D. □

Figure fig. 1 illustrates regions and points defined in lemma 2.1 for a simple function.

Figure 1: Illustation of regions from lemma 2.1



Theorem 2.2. Let C_l be the set of functions such 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 learning rate $\gamma < \frac{\mu^\dagger}{L^2_{\text{global}}}$ converges to $x^\dagger$ almost surely.
- Gradient descent initialized randomly in arbitrary 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. Proof is presented in [MJS23] in appendix E. □

Lemma 2.3. Consider gradient descent initialized randomly in set W with learning rate $\gamma \leq \frac{1}{2L}$. Let $X \subset \mathbb{R}^d$ arbitrary set of points in the landscape, f is L-smooth over $\mathbb{R}^d \setminus X$. Probability 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. Proof is presented in [MJS23] in appendix B. □

Theorem 2.4. Let X be an arbitrary 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 probability of not encountering any point of X during gradient descent with learning rate $\gamma \leq \frac{\mu_*}{L^2}$ 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. Proof is presented in [MJS23] in appendix C. □

3 Experiments

1D Example

Study different GD behaviour depending on starting point and learning rate on the function f.

$$f(x) := \begin{cases} -1600(x-2.5)^5 - 2000(x-2.5)^4 + 800(x-2.5)^3 + 1020(x-2.5)^2 & 2 \leq x \leq 3 \\ 1411.2 \times \left(1 - 10^4(x-8.4)\right) & 8.4 \leq x \leq 8.40001 \\ 0 & 8.40001 \leq x \leq 8.59999, \\ 1479.2 \times \left(10^4(x-8.6)+1\right) & 8.59999 \leq x \leq 8.6, \\ 20x^2 & \textit{otherwise} \end{cases}$$

Figure 2: Visual comparison of GD with different learning rates.

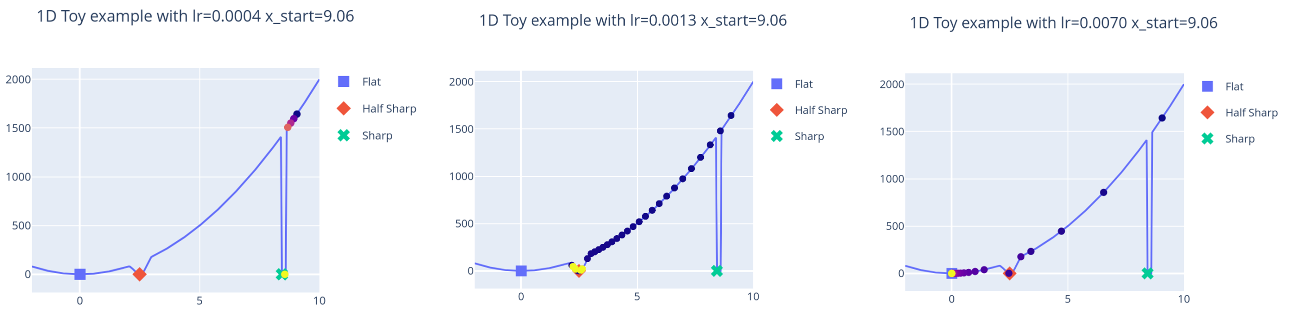
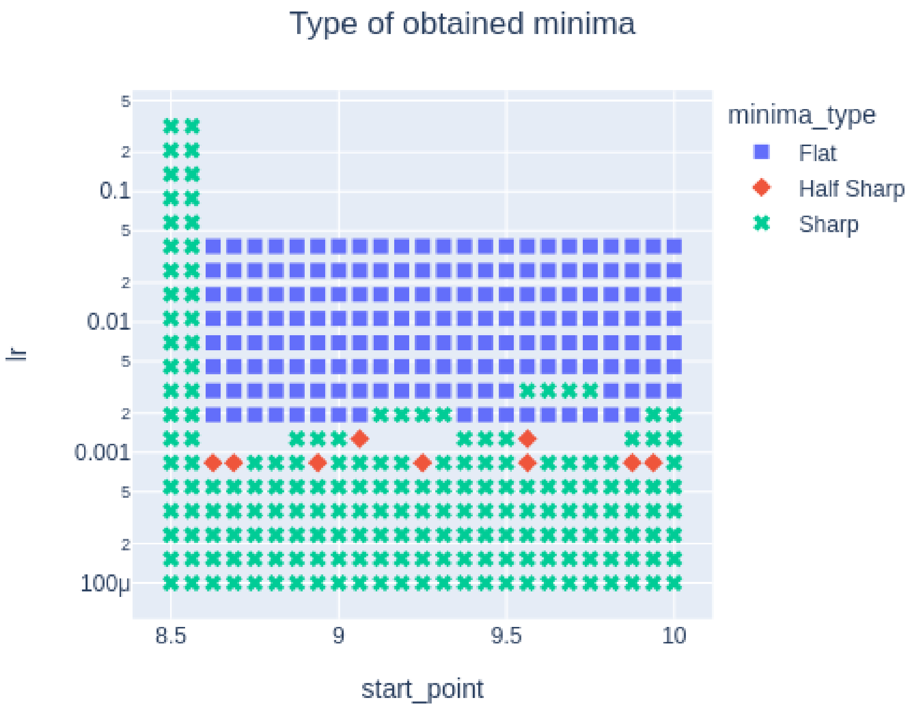


Figure 3: GD converged to the sharpest minima.

Figure 4: GD converged to the sharp minima.

Figure 5: GD converged to the flat minima.

Figure 6: Types of convegence of GD depending on startpoint and learning rate. Empty spaces represent no convergence in 1000 iterations.



From the experiments we can see that usage of larger learning rates helps to escape sharp minima and converge to a flatter one.

3.1 2D Example

Study different GD behaviour depending on learning rate on the function f

$$f(x,y) := x^2 + y^2 - 200\mathrm{ReLU}(|x|-1)\mathrm{ReLU}(|y|-1)\mathrm{ReLU}(2-|x|)\mathrm{ReLU}(2-|x|)$$

Figure 7: Visual comparison of GD with different learning rates.

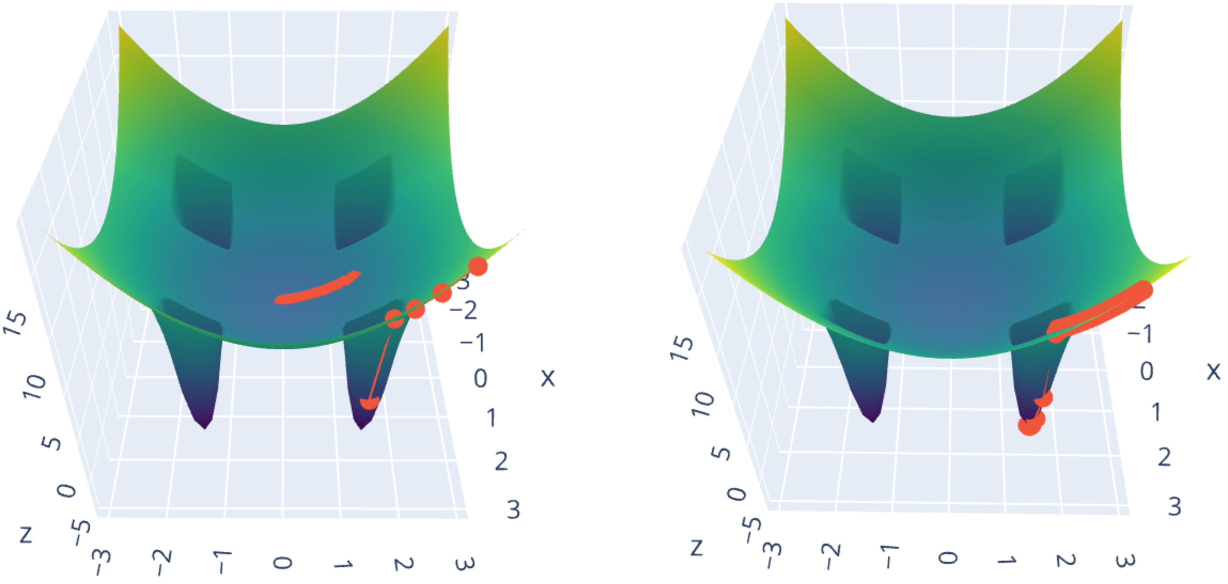
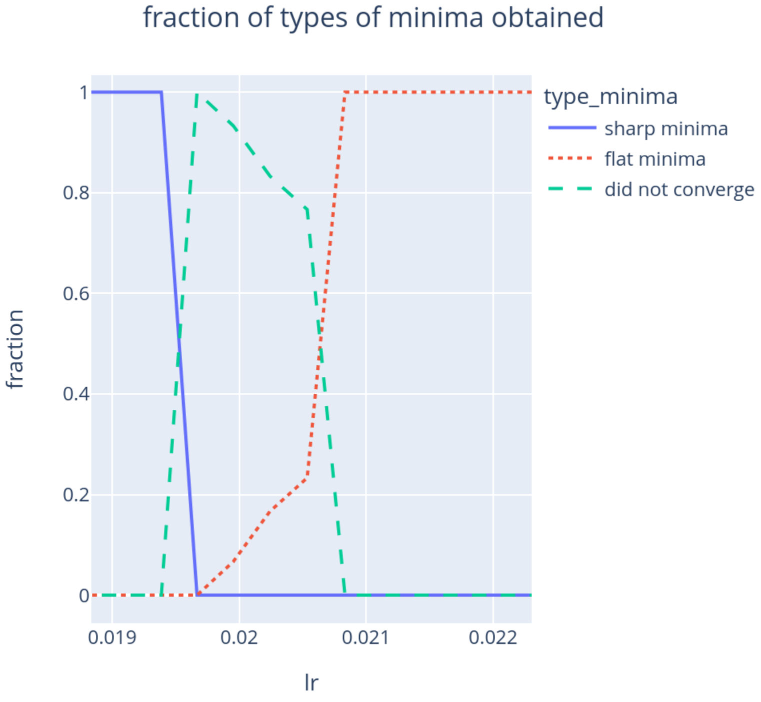


Figure 8: GD converged to flat minima.

Figure 9: GD converged to sharp minima.

Figure 10: Share of GDs that obtained flat minima if starting point is random $3 \leq x, y \leq 4$.

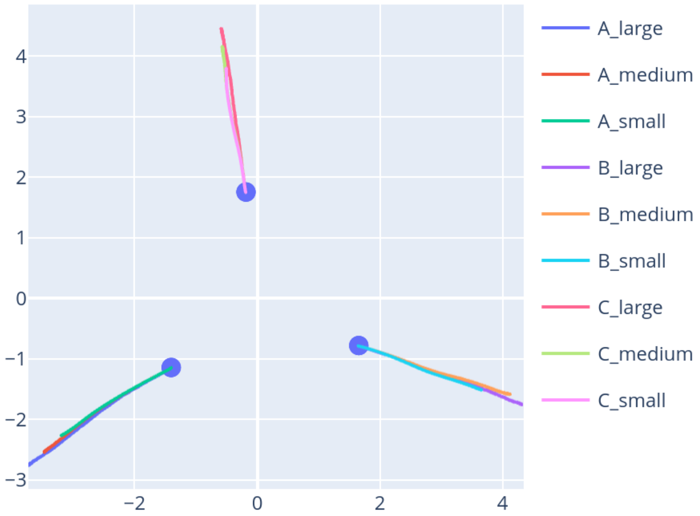


From these toy examples we can see that GD with larger learning rate indeed converge to flatter minima.

3.2 MNIST example

For the experiment with neural networks MNIST dataset was taken. The architecture of the NN consisted of 3 linear layers with ReLU activation layers each and cross entropy loss function. During the experiment 3 random initial points were chosen and from each starting point 3 GDs with different learning rates were performed. Weights of the first layer were saved during training and then were reduced to 2 dimensional space using PCA. PCA was performed in 2 ways. The first way is all starting points together, the second one is each starting point separately from the others.

Figure 11: Trajectories of all starting points



From this picture it seems like the trajectory does not depend much on a learning rate and GDs from same point follow similar path. But if we look at trajectories when PCA performed for each starting point independently we can see that the paths are different. Also some symmetry across different starting points can be noted.

Figure 12: Trajectories of each starting point.

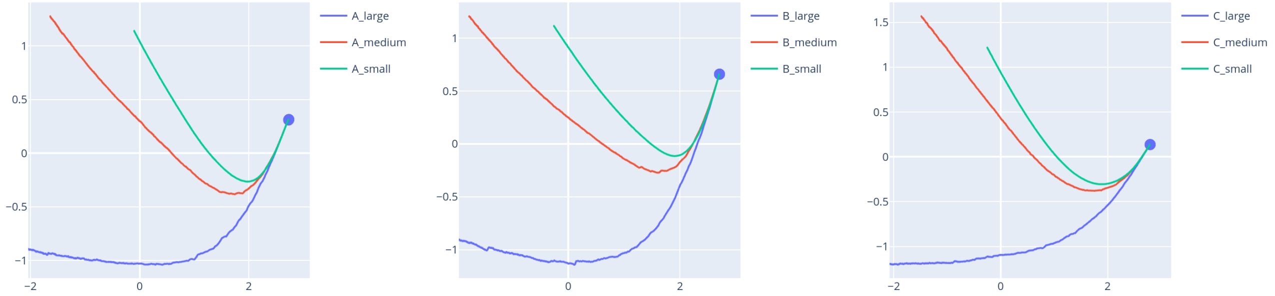


Figure 13: GD for starting point A. Figure 14: GD for starting point B. Figure 15: GD for starting point C.

All the obtained minimas showed similar performance on train and test validation.

4 Conclusion

It was shown that for certain class of functions and with weak assumptions on starting point GD with small learning rate almost surely converges to sharp minima, while GS with larger learning rate almost surely converges to flat minima. This behaviour was shown on one and two dimensional cases. Similar behaviour was shown for neural networks. For different starting points GDs with different learning rate converged to different minimas.

5 Future work

For more illustrative experiments a pretrained and more complex neural network might be taken. It is initially near some minima, and different learning rate can lead to converging to it, or escaping from it and search for a flatter minima. In this case obtained trajectories might have more notable difference. Also for more complex architectures flatter minima might be more robust and performe better on testing set.

References

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