Report on the article "Theory of Deep Learning IIb: Optimization Properties of SGD"

Team #13. Nick Osipov, Andrey Zharkov, Ann Vlasova, Ann Kuzmina

October 2018

1 Theoretical overview

The article researches the problem of characterization the convergence of stochastic gradient descent (SGD) on the non-convex empirical loss in deep learning. Authors conjecture that SGD, while minimizing the empirical loss also maximizes the volume, that is "flatness", of the minima. It was provided a few researches were this regularity was noticed. In [1] it was shown that neural networks training with mini-batches generalizes better than neural networks trained on large-batches. And it was noticed that stochastic gradient descent on mini-batches results into significantly flatter minima. It was observed that sharp minima lead to poor generalization, while flatten minima lead to good generalization.

The conjecture also has been tested for the case of deep networks of the compositional in authors' previous article [2].

The main idea is that if we have overparametrization (like in deep networks), than we have infinite set of solutions and this solutions correspond to flat regions during optimization.

It is consered gradient descent with white noise:

$$f_{t+1} = f_t - \gamma_t(\nabla I_{S_n}(f_t) + \xi_t)$$

where f – parameters of function (or weights of deep network), γ_t – learning rate, I_{S_n} – expected loss on training set S_n

It was shown (mainly empirical) that vector of gradient updates shows components with approximate Gaussian distributions (due to Central Limit Theorem since minibatch involves sum over many random choices of datapoints). And it is shown theoreticaly that the asymptotic distribution reached by SGDL – is the Boltzman distribution:

$$p(f) = \frac{1}{Z} e^{-\frac{U}{T}}$$

where Z is a normalization constant, U is the loss and T reflects the noise power. The equation implies that SGD prefers degenerate minima relative to

non-degenerate ones of the same depth. In addition, among two minimum basins of equal depth, the one with a larger volume, is much more likely in high dimensions.

2 Experiment results

In the article there are a few experiments that shows who SGD converges in cases of several attraction basins. Code is situated on this resource [3]

And there is an experiment demonstrating that good generalization adjacent with flat minima. It was taken CIFAR-10 and MNIST datasets and two models were fitted: one with true lables on training set, and one with random labels. It turned out that natural labels implies flat minima of loss function, while random labels implies sharp minima of loss function. And obviously model trained to natural labels has much better generalization ability than model trained on random labels.

2.1 SGDL on non-convex loss functions

In this part we explore behaviour of basic SGD method

$$f_{t+1} = \Pi_K(f_t - \gamma_t \nabla V(f_t, z_t)) \tag{1}$$

and his variation — SGDL

$$f_{t+1} = f_t - \gamma_t \nabla V(f_t, z_t) + \gamma'_t W_t \tag{2}$$

where W_t — Standart Gaussian vector and γ'_t — step-size.

At first, we construct non-convex loss function (3) with ,,flat“ and ,,sharp“ minimum (see Figure 1a)

$$f(x,y) = [0.1(x-4)^2 - 1]2^{-(x-4)^2} + [0.1(x-8)^2 + 0.1(y-5)^2]2^{-(x-8)^2-(y-5)^2} \tag{3}$$

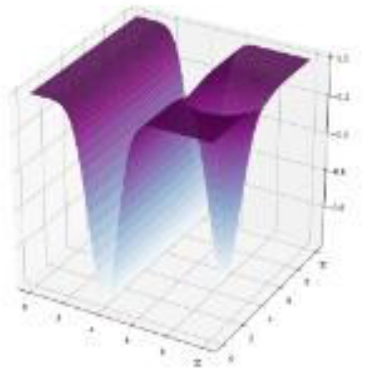
We run SGDL algorithm with different step γ'. Case with $\gamma' = 0$ is according to *SGD* algorithm.

Both methods show not the same results as on the paper. (Figure 2) Sharp minima is more ,,popular“ than the flat one for this algorithms whereas in the article the outcome is opposite.

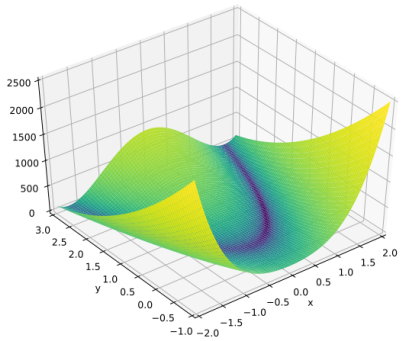
Moreover we try to run them on Rosenbrock function (4), which has global minima at $(x, y) = (1, 1)$ and $f(x, y) = 0$.

$$f(x,y) = (1-x)^2 + 100(y-x^2)^2 \tag{4}$$

In this case methods show bad results and and don't converge to minima (Figure 4)



(a) Constructed loss (3)



(b) Rosenbrock function

Figure 1: Non-convex loss functions

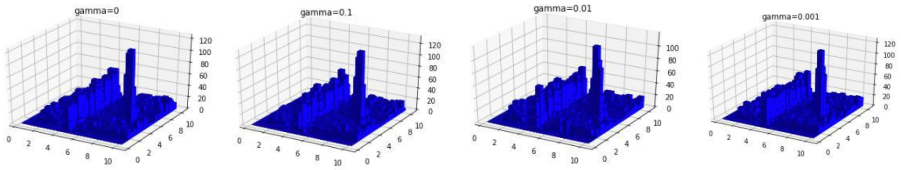


Figure 2: SGDL for loss function (3)

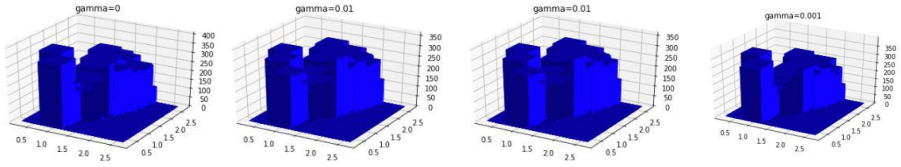


Figure 3: SGDL for Rosenbrock function

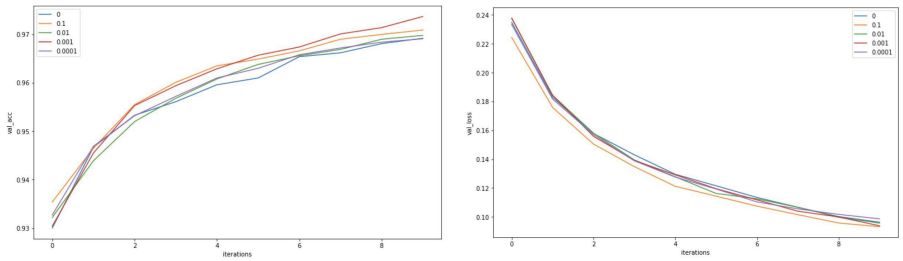


Figure 4: SGDL on Cifar-10

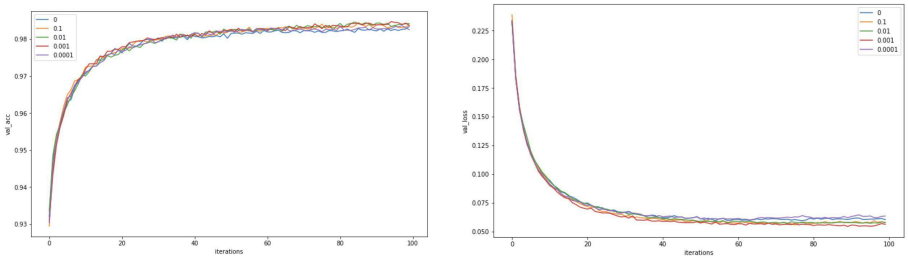


Figure 5: SGDL on MNIST

2.2 Convergence SGD on datasets

We use datasets MNIST and Cifar-10 to check convergence SGDL with different step-size γ'. All of them have similar good results (see Figure 4 and 5)

2.3 Random labels

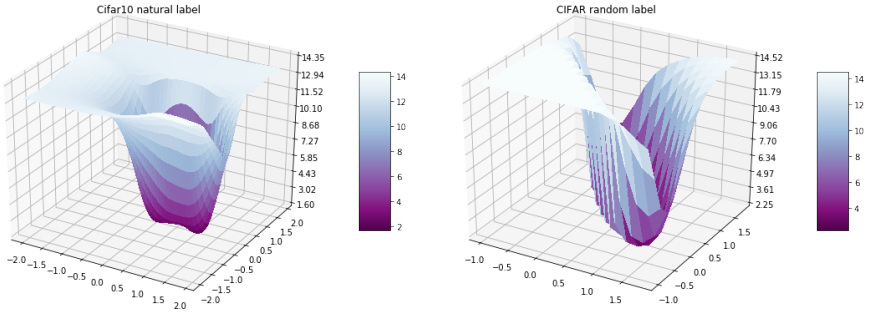
At the last part we check hypothesis that the flatness of the landscape around global minima of the empirical loss found by SGD, on CIFAR-10 and MNIST, with natural labels and random labels, respectively. Specifically, let ω_1, ω_2, ω_3 be three minimizers for the empirical loss found by SGD. For $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ on the simplex Δ_3, let

$$\omega_\lambda = \lambda_1\omega_1 + \lambda_2\omega_2 + \lambda_3\omega_3$$

For this purpose we plot "three-point interpolation". (see Figure 6 and 7). These results don't show any heavy flatness differentials as on the article.

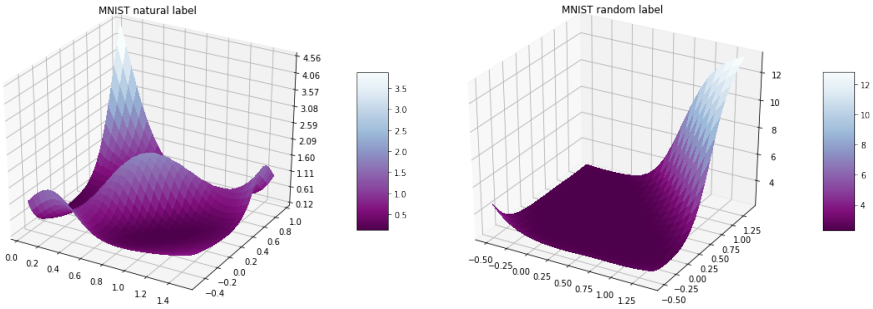
3 Conclusions

At first part of our research we get that SGDL converge to ,,sharp" minima more than to ,,flat" one (Figure 2). It happens probably because we use not the same functions that authors. The fact that algorithms don't work at Rosenbrock



(a) natural labels (b) random labels

Figure 6: Cifar-10



(a) natural labels (b) random labels

Figure 7: MNIST

function is expected because it has many local minimum and needed to use any heuristics except SGD.

Next we consider SGDL with different step-size on MNIST and Cifar-10 datasets and get the same results as in original paper.

Finally we explore behavior of empirical loss function on MNIST and Cifar-10 for natural and random labels. The random one have no sharp minima as the natural (Figure 7) that disprove claims on the original article.

References

[1] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836, 2016.

[2] Tomaso A. Poggio and Qianli Liao. Theory II: landscape of the empirical risk in deep learning. *CoRR*, abs/1703.09833, 2017.

[3] Andrew Zharkov. SGD exploration. `https://github.com/andreyzharkov/SGD_exploration`, 2018. [Online; accessed 02-November-2018].