Stochastic Gradient Descent in Over-Parameterized Regimes

Jeffrey Mei, Cody Melcher, Kamaljeet Singh
Department of Mathematics,
The University of Arizona,
jmei@math.arizona.edu, cmelcher@math.arizona.edu, kamaljeetsingh@math.arizona.edu

Abstract

As the success of large prediction models continues to grow, it becomes increasingly important to understand the geometry and behavior of these high dimensional parameter spaces. Recent evidence points towards a theory of model fitting that undercuts many classical results.

In this report, we provide a survey of these claims, and provide numerical experiments to evaluate some.

1 Introduction

The classical understanding of the bias-variance tradeoff was upended by the discovery of the double descent phenomenon [2]. Classical bias-variance tradeoff suggests that increasing model complexity inevitably leads to overfitting. A long standing mystery has been why neural networks have such successful generalization performance even when it has been "overfit" on training data. The double descent phenomenon demonstrates that the bias-variance tradeoff is a incomplete theory. For under-parameterized models, the bias-variance tradeoff successfully explains generalization performance. However, for high capacity models (e.g. neural networks, random forest, etc.), increasing model complexity often leads to improved generalization error – often better than the optimal suggested by the bias-variance tradeoff. Consequently, studying over-parameterized models – models that interpolate the training data to achieve perfect training error – have been an active field of research.

Recent evidence suggests SGD possesses many advantages in these over-parameterized contexts than in classical scenarios (Ma, et al., 2018; [1]). In particular, whereas SGD tends to reach local minima in classical scenarios, SGD tends to reach global minima in over-parameterized regimes.

In under-parameterized regimes, minibatch SGD scales linearly with the size of batches, i.e. the larger the minibatch size, the faster the convergence. However, in over-parameterized regimes the size of the minibatch has diminishing returns.

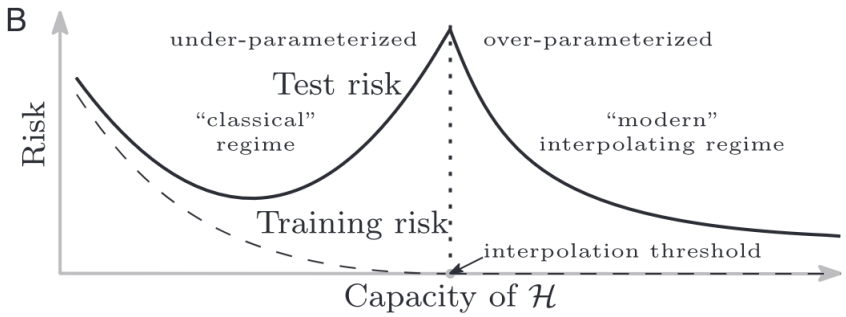


Figure 1

1.1 Applications

SGD is one of the most popular optimization procedures and has widespread applications. One of its primary advantages is that it yields significant computational savings. This is especially important as data and models grow increasingly large.

If the claim that the size of the minibatches saturate and yield diminishing returns, then we have gained an important insight into over-parameterized models.

1.2 Literature review

2 Methodology/Algorithm description

Describe the algorithms and state their convergence results.

Algorithm 1 Stochastic Gradient Descent

(0) **input:** $x_0 \in \mathbb{R}^m$
(1) **for** $k = 0 \dots T-1$ **do**
Take $I_k \subset \{1, ..., n\}$ s.t. $|I_k| = b << n$
$\alpha_k = 1/\sqrt{k}$
$x_{k+1} = x_k - \frac{\alpha_k}{b} \sum_{i \in I_k} \nabla f_i(x_k)$
(2) **end for**

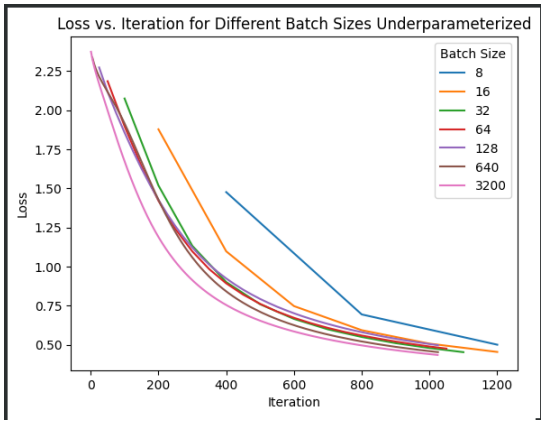
In Algorithm **??**
We impose the following requirements on f_i's...

Assumption 1. *For all...*

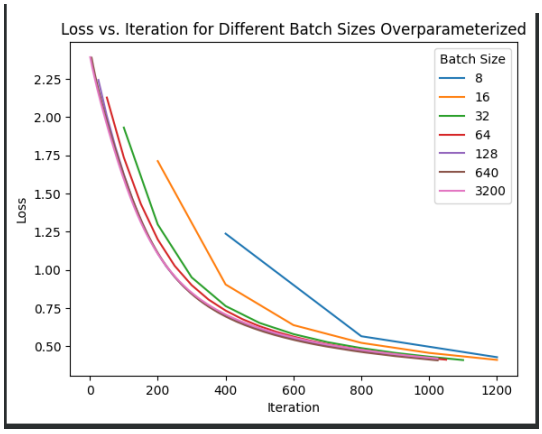
Theorem 1. *[?] Suppose Assumption 1 holds.*

3 Numerical Experiments

Conduct some numerical experiments to showcase the performance of the method(s) to solve the considered problem. Use Figures and Tables to show the performance of the methods.



(a) Under-parameterized loss per iteration



(b) Over-parameterized loss per iteration

Figure 2: Numerical results for batch-size saturation

3.1 Batch Size Saturation

This section of the project investigates the convergence behavior of Stochastic Gradient Descent (SGD) in an overparameterized model. According to Section 4.4 of [1], two distinct regions exist based on the batch size, denoted by a critical batch size m^*, where:

1. When the batch size is less than m^*, one iteration of SGD with a mini-batch size m (where $m \leq m^*$) is equivalent to m iterations of mini-batches of size 1, up to a multiplicative constant. (Linear) 2. When the batch size is greater than m^*, one iteration of SGD is equivalent to either one iteration of m^* or one iteration of full gradient descent, up to a multiplicative constant. (Saturation)

To validate this claim, we examined the convergence behavior of our model in both under-parameterized and over-parameterized settings. We trained our model using varying batch sizes. The hypothesis is that after reaching a specific batch size (approximately m^*), there should be diminishing returns for further increases in batch sizes.

Implementation: We trained two neural networks, each with a single hidden layer. The size of the hidden layer is set to 32 for the under-parameterized model and 128 for the over-parameterized model. We experimented with various batch sizes, including 8, 16, 32, 64, 128, and 640, along with the number of training examples. We extracted the training error at the end of each iteration from the model's training history.

The results clearly indicate that in the over-parameterized model, the curves converge and become almost overlapping after a certain batch size (32), suggesting saturation. This suggests that the critical batch size m^* is approximately 32 for the over-parameterized model.

However, in the under-parameterized model, the curves do not merge into each other at or after any batch size. There is a clear spectrum of curves for varying batch sizes, indicating the absence of Linear Scaling and Saturation regions in the under-parameterized regime.

3.2 Local Minima is Global Minima

The literature we have reviewed has often made the claim that minima reached in overparameterized regimes are increasingly likely to be global. However, while there is a lot of theoretical analysis on the subject that support this claim, empirical evidence remains scarce. We propose an elementary

test to probe this claim.

If local minima are common in under-parameterized regimes, we can expect that if we randomly initialize a neural network multiple times, the generalization error should

In short, we compare the variation in the

4 Conclusion and Future Direction

Since the discovery of the double-descent phenomenon, a wide frontier of research has opened up.

Given more time, we would like to investigate the claim that in over-parameterized regimes, local minima have a tendency to also be global minima.

References

[1] M. Belkin, *Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation*, Acta Numerica, 30 (2021), pp. 203–248.

[2] M. Belkin, S. Ma, and S. Mandal, *To understand deep learning we need to understand kernel learning*, June 2018. arXiv:1802.01396 [cs, stat].