Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Bachelor Thesis Cognitive Science

# Investigating Probabilistic Preconditioning on Artificial Neural Networks

Ludwig Valentin Bald

October 14, 2019

**Gutachter**

Prof. Dr. Philipp Hennig
(Methoden des Maschinellen Lernens)
Wilhelm-Schickard-Institut für Informatik
Universität Tübingen

**Betreuer**

Filip De Roos
(Methoden des Maschinellen Lernens)
Wilhelm-Schickard-Institut für Informatik
Universität Tübingen

**Bald, Ludwig:**
*Investigating Probabilistic Preconditioning on Artificial Neural Networks*
Bachelor Thesis Cognitive Science
Eberhard Karls Universität Tübingen
Thesis period: 19.06.2019-18.10.2019

# Abstract

In Deep Learning, many different optimization algorithms are used. Second-order methods use curvature information to take good optimization steps, but are computationally very intensive. First-order methods only have access to the gradient and are susceptible to ill-conditioned problems, whose loss landscape curves much more strongly in one direction than in others.

The performance of first-order methods can be improved by preconditioning the problem, which transforms the loss landscape to curve more equally in all directions. In this thesis, I present and investigate an implementation of the probabilistic preconditioning algorithm proposed in [Roos and Hennig, 2019] using the optimizer benchmarking framework DeepOBS, as proposed in [Schneider et al., 2019].

I present evidence that suggests the algorithm does in fact improve the learning speed of Stochastic Gradient Descent. However, it does not outperform other adaptive methods like Momentum and comes with additional computational overhead.

# Zusammenfassung

Im Feld von Deep Learning wird eine Reihe an Optimierungsalgorithmen verwendet. Methoden zweiter Ordnung benutzen Information über die Krümmung der Fehlerlandschaft eines Problems um sehr gute Optimierungsschritte zu finden, sind aber sehr rechenaufwändig. Methoden erster Ordnung benutzen nur den Gradienten der Fehlerfunktion und sind anfällig für schlecht konditionierte Probleme, also solche, deren Fehlerlandschaft in verschiedene Richtungen verschieden stark gekrümmt ist.

Die Ergebnisse von Methoden erster Ordung können durch Preconditioning verbessert werden. Dabei wird die Fehlerlandschaft so transformiert, dass sie in alle Richtungen ähnlicher gekrümmt ist. In dieser Arbeit zeige und untersuche ich eine Implementierung des probabilistischen Preconditioning-Algorithmus, vorgestellt in [Roos and Hennig, 2019]. Dabei verwende ich das Optimierer-Vergleichs-Framework DeepOBS, vorgestellt in [Schneider et al., 2019].

Ich präsentiere Evidenz dafür, dass der Algorithmus tatsächlich die Ergebnisse von Stochastic Gradient Descent verbessert. Jedoch bleibt er hinter anderen adaptiven Methoden wie Momentum zurück. Dabei verwendet der Preconditioner zusätzliche Ressourcen.

# Acknowledgments

I would like to thank Aaron Bahde and Frank Schneider for developing the excellent benchmarking framework DeepOBS, which was essential to this thesis, and for always being quick to answer questions and fix bugs.

I would also like to thank my supervisor Filip De Roos, who was always available and helpful whenever I had questions.

# Contents


**1 Introduction** **1**

**2 Fundamentals and Related Work** **3**

2.1 Probability Basics . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2.2 The Normal Distribution . . . . . . . . . . . . . . . . . . . . . . . . 4

2.3 Machine Learning . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.3.1 Loss functions . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.4 Optimization . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

2.4.1 Gradient Descent . . . . . . . . . . . . . . . . . . . . . . . 6

2.4.2 Stochastic Gradient Descent . . . . . . . . . . . . . . . . . 7

2.4.3 Second-Order Methods . . . . . . . . . . . . . . . . . . . . 7

2.5 Preconditioning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.6 Deep learning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

2.6.1 Artificial Neural Networks . . . . . . . . . . . . . . . . . . 9

2.6.2 Automatic Differentiation . . . . . . . . . . . . . . . . . . 10

2.7 Benchmarking . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

**3 Approach and Implementation** **11**

3.1 Description of the algorithm . . . . . . . . . . . . . . . . . . . . . . 11

3.1.1 Modifications of the algorithm . . . . . . . . . . . . . . . . 12

3.2 Documentation for the class Preconditioner . . . . . . . . . . . . . 13

3.2.1 Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

3.2.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

3.2.3 Implementation Details . . . . . 14
3.3 Test Problems . . . . . 16
3.4 DeepOBS baselines . . . . . 17
3.5 Technical details . . . . . 17

**4 Experiments** **18**
4.1 Experiment 1: Preconditioning . . . . . 18
4.2 Experiment 2: Wallclock Time . . . . . 19
4.3 Experiment 3: Initialization . . . . . 22
4.4 Experiment 4: Learning Rate Sensitivity . . . . . 23
4.5 Discussion . . . . . 25
4.6 Further Research and Development . . . . . 28
4.7 Feedback for DeepOBS . . . . . 28

**5 Conclusion** **30**

**A Code** **31**
A.1 Singularity build recipe . . . . . 31
A.2 DeepOBS Runscript for experiment 1 . . . . . 32
A.3 Slurm Batch Definition File . . . . . 33
A.4 Repository . . . . . 34

**References** **34**

# Chapter 1

# Introduction

In recent years, machine learning and more specifically Deep Learning has been becoming more and more relevant. Both in research and in application, it is ubiquitous. It has lead to anything from a better understanding of the brain's structure ([Brown and Hamarneh, 2016]) to major advancements in self-driving cars.

Even though the practical use of the current state of machine learning is unquestionable, there is a lot of opportunity for research and further improvement of the inner workings. One of the areas of interest is the study of optimization algorithms. They have a large effect on the time it takes to find an optimal parametrization while training a machine learning model. Improvements in this area lead to reductions in cost, need for data and energy consumption.

The simplest optimizer is Stochastic Gradient Descent (SGD), which is a first-order-method and needs access to the gradient of the error function of the model. Second-order methods make use of the Hessian of the error function, which is the second derivative and contains information about curvature. The additional curvature information allows these methods to take much larger and more precise steps, but they require more computational effort which is infeasible for high-dimensional problems like Deep Learning models.

Two recent publications mainly drive this thesis: [Roos and Hennig, 2019] describes a preconditioning algorithm that estimates the Hessian and applies its inverse to the parameters, which is a way of including second-order information. This changes the parameter landscape so that standard first-order optimizers can converge faster to the optimum. The paper suggest that it compares favourably in performance over non-preconditioned, standard SGD. The paper also reports results for a single Deep Learning experiment. The second paper, [Schneider et al., 2019], presents the benchmarking suite "DeepOBS" which makes it easier to compare new optimization algorithms against established baselines in an unbiased way.

This thesis makes multiple key contributions. The first is a reimplementation of the preconditioning algorithm, which increases usability and adds functionality, like the possibility to easily switch out the inner optimizer which takes over after preconditioning. The second contribution is the comparison of the preconditioning algorithm against established benchmarks measured using DeepOBS, which includes both an attempt to reproduce the results and further investigation into applying the optimizer to Deep Learning models. In parallel with this thesis, a pyTorch version of DeepOBS was being developed, which is described in detail in [Bahde, 2019]. Another contribution was to provide user feedback to the development team of DeepOBS. A selection of main feedback points is replicated towards the end of this thesis.

# Chapter 2

# Fundamentals and Related Work

The studied algorithm builds on concepts from probability theory, linear algebra and numerical optimization. This chapter also contains a high-level introduction to Deep Learning.

The following concepts in this chapter are adapted from [Bishop, 2006], if not otherwise specified.

## 2.1 Probability Basics

This thesis makes use of some basic concepts of probability theory. Most notably Bayes' Rule, point estimates, and vector-valued normal distributions.

Bayes' Rule tells us what happens to the probability of an event $X$ after observing new evidence $E$, given a *prior* probability $P(X)$ and probabilities $P(E)$ and $P(E|X)$. It is given by this formula:

$$P(X|E) = \frac{P(X) \cdot P(E|X)}{P(E)}$$

$P(X|E)$ is the *posterior* probability.

When modelling a random variable $x$, a point estimator $\hat{x}$ is a function that gives an approximation for the true value for $x$ based on evidence, for example observations of $x$. One example of a point estimator is the $maximum - a -$ *posteriori* estimate, which is defined as precisely the value for $x$ which has the highest posterior probability, i.e. the maximum of the posterior probability distribution.

## 2.2 The Normal Distribution

One of the central concepts in probability theory is the Gaussian distribution. The real-valued *normal* or *Gaussian* distribution is defined as:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

It has the parameters *mean* $\mu$ and *variance* $\sigma^2$. The mean represents the maximum-a-posteriori estimate for $x$; it is the maximum of $N(x|\mu,\sigma^2)$ Extending this to a *multivariate* Gaussian distribution with $n$ dimensions, the definition needs to be adapted:

$$\mathcal{N}(\vec{x}|\vec{\mu},\Sigma) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right)$$

where $\vec{x}$ or simply $x$ is the $n$-dimensional vector of random variables. The parameters are similar to the one-dimensional case. The $n$-dimensional vector $\vec{\mu}$ or simply $\mu$ denotes the mean. The symmetric $n \times n$ matrix $\Sigma$ is the *covariance matrix*, where

$$\Sigma_{ij} = \text{cov}(x_i, x_j) = \Sigma_{ji}$$

For the diagonal entries it follows that:

$$\Sigma_{ii} = \text{cov}(x_i, x_i) = \text{var}(x_i)$$

If the variables $x_i$ are independent, $\Sigma$ is a diagonal matrix.

Matrix-valued normal distributions can be written as vector-valued normal distributions by flattening the matrix to a single vector. Bayes' rule can also be applied to vector-valued normal distributions, which again leads to vector-valued normal distributions as posteriors.

## 2.3 Machine Learning

Machine Learning algorithms are central to almost every area of modern society. They control which products we buy, which music we discover and they enable recent advances in industrial automation and autonomous driving. Specialist Artificial Intelligence agents routinely outperform the best human players in the hardest games, like Go ([Gibney, 2016]) or Starcraft II ([Vinyals et al., 2019]). This recent rise of machine learning can be attributed to a number of factors. Enabled by the internet, almost every interaction with a technological system is tracked and recorded. Large, high-quality data sets are available to everyone with an internet connection through Open Data initiatives. Learning algorithms have been around for a few decades, but recent

advances in hardware meant it became economically feasable to use them on a large scale.

The general process of machine learning can be summarized in multiple steps:

1. Gathering Data: The gathered data needs to be prepared and split into multiple sets, one for training, one set for testing generalization and a third set for validation of hyperparameter tuning.
2. Chosing the model: Chosing a suitable model depends on the problem, requirements and other constraints. A possible choice for image classification tasks is for example a Convolutional Deep Neural Network.
3. Training the model: The model is shown the data, and its parameters are adjusted towards the optimal setting, as measured by a predetermined metric. The process is often run multiple times with different optimizer hyperparameters to find the best setting, which is called *hyperparameter tuning*.
4. Application: The model is tested on the previously prepared test data set to see how it performs. If it does well, it can be deployed on the intended task.

### 2.3.1 Loss functions

In order to find an optimal solution, there needs to be a measure of "goodness" of a given parametrization. The loss function usually is a sort of distance measure between the model's outputs and the true labels of the training data.

$$L : \mathbb{R}^n \to \mathbb{R}$$

One example for a loss function for classification problems, which is used for the experiments in this thesis, is Cross Entropy Loss, where $w$ are the model's parameters, $y$ is the true label, $\hat{y}_w$ is the model's output.

$$L(w) = -\sum_{1}^{K}(y \cdot \log(\hat{y}_w(x)))$$

A model can achieve very good scores on training data, but fail to generalize these results to unseen data. This unwanted behavior is called *overfitting*. In order to prevent the model from overfitting, often a regularization term is added:

$$L(w) = -\sum_{1}^{K}(y \cdot \log(\hat{y}_w(x))) + \lambda\|w\|^2$$

Here $\lambda$ is a parameter that determines the strength of the regularization. This will penalize large weights during training, which biases the model towards

less extreme parametrizations which are caused by fitting too closely to training data. In most cases L2 regularization is equivalent to weight decay, or decreasing weights after each optimization step. See [Chaudhari et al., 2017] for futher investigations on the differences between these two.

It's worth noting that the choice of loss function and the regularization term have a significant impact on the outcome and convergence speed of training. Ultimately, other measures like classification accuracy might be more important during application.

## 2.4 Optimization

In order to find the optimal values for the parameters of a model, different methods have been proposed. For low-dimensional optimization problems with a small number of parameters, it is often possible to find the optimal parametrization analytically. For high-dimensional optimization problems, the analytical solution often is computationally intractable.

Numerical algorithms address this problem by using the available local information in order to iteratively approximate the globally optimal parametrization. The number of iterations needed until the algorithm converges varies based on the implicit prior assumptions about the model. Some of these optimization processes expose *hyperparameters*, which influence the performance of the algorithm. These are different from the model's parameters, which we want to optimize.

Optimization algorithms can be grouped by the order of the highest-order derivative they use. There are zeroth-order methods that only require the value of the loss function itself. First-order methods like Gradient Descent use the first derivative, or the gradient. They work under the assumption that the gradient points in the direction of the global minimum and approaches zero while approaching the minimum.

### 2.4.1 Gradient Descent

If the gradient of the Loss function at a certain point in parameter space is known, this information can be used to update the parameters in the direction towards the solution. A widely used family of optimization algorithms is derived from Gradient Descent. Gradient Descent means: Just take a step in the direction of the steepest gradient. Scale the step size by the steepness of the gradient. If the gradient is very steep, take a larger step. If it is small, take a smaller step, like a drunk student taking repeated tumbles down a hill and ending up on a local minimum.

$$w_{i+1} = w_i - \alpha \cdot \nabla L(w_i)$$

Gradient Descent only has a single hyperparameter, the "learning rate" $\alpha$, which scales the step. Its optimal choice depends on the problem and is generally not obvious.

### 2.4.2 Stochastic Gradient Descent

Traditional Gradient Descent is a deterministic model, which in Deep Learning is equivalent to using the whole data set to compute the true gradient. Given a large data set, one can better use their computational ressources by taking several noisy smaller steps instead of one precise step. This variation is called *Stochastic* Gradient Descent. Instead of the true loss, which is the model's performance on the whole training data, it computes only an estimate for the true loss by using a subset or *Minibatch* of the training data. A common choice is between 32 and 256 data items per minibatch. This greatly improves convergence speed, as the required computations are much easier to perform. However, especially for smaller batch sizes, this adds noise to the system, meaning that the parameter update step points only roughly in the direction of the steepest actual gradient.

$$w_{i+1} = w_i - \alpha \cdot \nabla \hat{L}(w_i)$$

Many variants of SGD have been proposed, for example by adding a momentum term or otherwise dynamically adapting the learning rate.

### 2.4.3 Second-Order Methods

First-order methods scale their step length by the gradient of the Loss function. This means that they get stuck on flat plateaus, because the gradient is very small. Adaptive Modifications of SGD like Adam and Momentum keep track of previously seen gradients in order to take better steps. This implicitly assumes that past gradients contain information about future gradients, which is often the case. For example, if there is little change in the last observed gradients, one can assume that the optimizer is far away from the minimum where gradients should converge to zero. Therefore, it is now able to take larger steps.

Second-Order-Methods are methods that explicitly use the Hessian $B$ of the Loss function, which is equivalent to the second derivative$\nabla\nabla L(w_i)$. This means they can use the explicit representation of curvature to take even better optimization steps.

Newton's method is the theoretical base for Quasi-Newton methods that try to achieve similar performance while limiting computational complexity.

Its update rule looks like this:

$$w_{i+1} = w_i - \frac{\nabla \hat{L}(w_i)}{\nabla\nabla \hat{L}(w_i)}$$

Note the similarity to SGD's update rule, but replacing the scalar learning rate with a matrix: $\alpha \equiv 1/\nabla\nabla \hat{L}(w_i)$

## 2.5 Preconditioning

SGD takes large steps if the gradient of the loss landscape is steep, and it takes small steps if the gradient of the loss landscape is flat. The underlying assumption is that as the gradient gets flatter (approaches zero), the algorithm approaches the minimum of the loss function. Taking large steps close to the minimum most likely means taking a step away from the minimum. However, SGD is not aware of the scale of the loss landscape. Some problems are generally flatter and have low general curvature than others. On these generally flat problems, flat gradients can be observed quite far from the minimum and mandate larger steps. Recalling the update function of SGD, the learning rate parameter $\alpha$ is a scaling factor for the gradient, which can correct for non-optimal general steepness:

$$w_{i+1} = w_i - \alpha \cdot \nabla \hat{L}(w_i)$$

Tuning the learning rate amounts to finding a global scaling factor for the gradient.

Often, machine learning problems are scaled differently in different directions. In one direction with low curvature, the problem might be quite flat, while in another direction with high curvature, gradients are generally steeper. This means that the optimal step length depends not only on the steepness of the gradient, but also its direction (In the low curvature direction, longer steps are warranted than in the high curvature direction). It is impossible to fully capture and correct for this structure using a simple scalar learning rate. Chosing the appropriate small learning rate for the high curvature direction will make the algorithm learn very slowly in the low curvature direction. Chosing anything larger than that runs a risk of exploding gradients in the high curvature direction, as the steps are too large. Using a matrix $P$ (the *Preconditioner*[1]) to scale every entry of the gradient matrix individually, we get the following update rule for Preconditioned SGD:

$$w_{i+1} = w_i - \alpha \cdot P \cdot \nabla \hat{L}(w_i)$$

[1]Technically, a Preconditioner is any transformation which reduces the *condition number*, or the ratio between largest and smallest eigenvalues of a problem.

A perfect preconditioner would be the inverse of the true Hessian of the loss function, which exactly captures the curvature of every parameter. If it were available, the true Hessian could be used for second-order methods, eliminating the need for preconditioning. Computing the true Hessian is very hard and expensive in the high-dimensional and noisy deep learning setting, so various approximations have been studied (see [Saad, 2003] for a detailed, more rigorous description)

## 2.6 Deep learning

### 2.6.1 Artificial Neural Networks

A popular machine learning paradigma is living through a resurgence: Artificial Neural Networks. The fundamental building block is the single neuron, which somewhat resembles a biological neuron. Its activation depends on the sum of the activation of its inputs. A neural network model is a connected sequence of neuron layers. There is an input layer which is a direct mapping of the training data point. The input layer' activation are fed forward into the "hidden layers", which in turn feed their activations to the neurons of the output layer. Information about observed data is stored in the model's parameters, the weights and biases of each layer. As a mathematical object, a neuron is an activation function which depends on the sum of the weighted activations of the inputs. Often there is a bias, which is a static value added to the input activation. Activation functions are not necessarily linear. For example, the Rectified Linear Unit is widely used:

$$ReLU(x) = \max(0, x)$$

The full activation of a Neuron $x_j$ in layer $\ell$ in a model with the layer's input weight matrix $W^{\ell-1}$ and bias $b_j$ is given by:

$$x_j^\ell = \sigma\left(\sum_i W_{ji}^{\ell-1} x_i^{\ell-1} + b_j\right)$$

$W_{ji}^{\ell-1}$ refers to the weight between neuron $i$ in layer $\ell - 1$ and neuron $j$ in layer $\ell$.

Different architectures of neural networks have been studied, which are useful for different tasks and types of input data. For example, fully connected Convolutional Neural Networks (CNN) are good at encoding visual or spatial information, while Recurrent Neural Networks have an internal state and are good at encoding information over time.

### 2.6.2 Automatic Differentiation

First- and higher-order optimizers require information about the derivatives $\nabla^n L$ of the loss function $L$ at a given point in parameter-space.In neural networks, this is achieved by an algorithm called *backpropagation*. Loss functions in neural networks are the composition of all the layerwise activation functions which are all individually differentiable. This structure leads to the algorithm of backpropagation, which computes the gradient by doing a *forward pass* followed by a *backward pass*.

In the forward pass, the loss function for a given minibatch is computed by applying the model to the minibatch data and computing the error score on its output. During these computations all operations on the hidden layers are recorded in a graph structure. For each of these basic operations computing the gradient is trivial. Taking the final loss value as the root of the graph, the algorithm traverses backwards through the graph and computes individual partial derivatives for each parameter by applying the chain rule along the path.

## 2.7 Benchmarking

There are no standard established benchmarking protocols in research on optimizers. It is unclear what measures to consider, or how they are to be measured. This means there is no rigorous base for choosing an optimizer as a practitioner. DeepOBS is a solution to this problem, standardizing a protocol, providing benchmarks and standard test problems.

It includes the most used standard datasets and a variety of neural network models to train as standard problems. Also, the performance of a variety of widely-used optimizers comes with DeepOBS as a baseline, which means these runs do not have to be repeated by every single researcher. DeepOBS specifies a protocol for hyperparameter tuning(see [Bahde, 2019]). Otherwise, researchers with ressources for extensive hyperparameter tuning would have an edge over those who can not afford such extensive tuning. It also provides the software suite to run the actual experiments on optimizers. It takes care of logging parameters and evaluating success measures. The analyzer class can generate matplotlib plots showing the experiment data.

All in all, DeepOBS fixes most parameters that need to be fixed when evaluating optimizers, which provides for a reasonably solid foundation for comparison.

# Chapter 3

# Approach and Implementation

As explained in the previous chapter, preconditioning is a way to modify a problem in order to improve convergence speed of first-order optimization methods. Constructing a computationally tractable preconditioner while using only noisy observations of Hessian-vector products was the goal of the recent paper by de Roos and Hennig. In this chapter, I give a short overview of the algorithm and explain my changes and the details of the implementation and experiment setup.

## 3.1 Description of the algorithm

As described in section 2.5, a perfect preconditioner would be the inverse of the Hessian of the loss function, wich is not directly accessible and expensive to compute. The algorithm tested in this thesis is described in detail in [Roos and Hennig, 2019]. This is a high-level overview on the algorithm's structure. The algorithm is made up of four main parts, which will be tested seperately throughout this thesis. It restarts by default every epoch.

1. Using observations of Hessian-vector-products, construct a multivariate Gaussian distribution as a prior estimate for the Hessian. Using this information about curvature, calculate a scalar measure $\alpha$ for the inverse of the curvature and use it as the learning rate of the inner optimizer.
2. Gather more observations and calculate the posterior multivariate Gaussian distribution using Bayes' rule and the prior constructed in step 1.
3. Take the mean of this distribution as the maximum-a-posteriori-estimate. Invert it.
4. For every following minibatch, rescale the observed gradient by applying the preconditioner, and perform an update step of the inner optimizer.

### 3.1.1 Modifications of the algorithm

The implementation and theoretical work lead to the following proposed changes to the abstract algorithm.

#### Parameter groups

Mainly due to technical reasons (see section 3.2.3), support was added for parameter groups, but this also yields an interesting theoretical change. The algorithm already treated every parameter layer as an independent task while inverting the Hessian. It however estimated a global step size, which was the same for every parameter. With this modification, the user can specify which parameters should be grouped together and get a common learning rate.

As described in section 2.5, the learning rate corrects for the scale of the overall curvature of the loss landscape. However, depending on the problem, different parameters require different scale factors. With this modification, scale factors are neither shared for all factors (as in SGD) nor down to an individual parameter precision (Preconditioning). Rather than going for the safest choice, the modified algorithm is able to use larger step sizes if all parameters in a group allow for it. This accelerates learning in the direction of those parameters. The benefit of this approach in practice was not tested in this thesis.

#### Automatic Assessment of the Hessian's quality

The algorithm estimates a low-rank estimate for the Hessian. After a number of optimization steps, the surrounding loss landscape has changed and the estimate for the Hessian no longer is useful. It is not clear when this re-estimation should optimally happen. An answer needs to balance the added computational cost of re-estimating the Hessian with the performance benefit that comes from having a better-fitting Hessian. Using the original algorithm, the estimation process is restarted every epoch. The authors do not justify this practical choice. The `Preconditioner` class includes a method `maybe_start_estimate()` which is called before every step of the optimizer and could be used to dynamically assess whether the Hessian is out of date.

Assessment of the Hessian's quality is made especially difficult by sampling noise. A possible approach is to test prediction power of the Hessian. Given the last observed gradient, how well does the Hessian predict the gradient for the next minibatch? Standard statistical techniques like a $\chi^2$ homogenity test could be applied, varying the confidence level to find the balance of having a good preconditioner while limiting the number of times it is computed. If the algorithm kept track of a measure of uncertainty for the Hessian and in turn its predictions, Bayesian methods would be possible to apply.

**Focusing on the adaptive learning rate**

Deep learning problems are very high-dimensional. [Chaudhari et al., 2017] report that around 10% of eigenvalues are large, while 90% are close to zero. A proper preconditioner would need to rescale around 10% of parameters. This means the proper preconditioner's rank would still be very high. Computations would be much more expensive and would probably eat up the performance benefit that comes with preconditioning. In experiment 1 it is tested whether applying the low-rank approximate preconditioner has an effect on convergence speed.

## 3.2 Documentation for the class Preconditioner

### 3.2.1 Overview

The class `Preconditioner` provides an easy way to use the probabilistic preconditioning algorithm proposed by [Roos and Hennig, 2019]. It's written in python and made to work with pyTorch. This is how to use it:

1. Get the source file from the public repository and include it in your project.
2. Initialize the preconditioner like any other optimizer. There are reasonable default values for the hyperparameters.
3. Depending on the version you're using, manually call `start_estimate()` at the beginning of each epoch.

In the next section there is more detailed documentation for the class, its attributes and functions.

### 3.2.2 Methods

- Public functions
  - `Preconditioner(params, est_rank=2, num_observations=5, prior_iterations=10, weight_decay=0, lr=None, optim_class=torch.optim.SGD, **optim_hyperparams)`
    - `params`: The model's parameters that will be optimized
    - `est_rank` : An Integer, the rank of the preconditioner
    - `num_observations` : An Integer, the number of posterior updates to the estimated Hessian
    - `prior_iterations`: An Integer, the number of iterations to construct the prior
    - `weight_decay` : A float between 0 and 1, the amount of weight decay

        - `lr`: The learning rate. If specified, it will be passed to the inner optimizer as `lr` in the first epoch.
        - `optim_class` : The `torch.optim.optimizer` class to be used as inner optimizer.
        - `**optim_hyperparameters` : Any other hyperparameters to be used for the inner optimizer
    - `start_estimate()`
        - (Re)starts the process of estimating the Hessian.
    - `step()`
    - `get_log()`
    - (`maybe_start_estimate()`)
- Private functions
    - `_initialize_lists()`
    - `_init_the_optimizer()`
    - `_gather_curvature_information()`
    - `_estimate_prior()`
    - `_setup_estimated_hessian()`
    - `_apply_estimated_inverse()`
    - `_hessian_vector_product()`
    - `_update_estimated_hessian()`
    - `_create_low_rank()`
    - `_apply_preconditioner()`

### 3.2.3 Implementation Details

In the following, I will highlight and explain some key software design decisions I took while implementing and refactoring the algorithm. The starting point was the original code from [Roos and Hennig, 2019]

The **main goals** for the implementation were to make the algorithm as easy to use as possible for a standard usecase, while maintaining the flexibility I needed to research specific variations. The conceptual modifications to the algorithm are discussed in section 3.1.1. The simple, default usecase was a user trying to use the preconditioner as proposed in the original paper, to optimize a neural network model. In this case, the necessary changes to the user's existing code should be minimal, with hyperparameters set to reasonable default values. For development, it is important to understand the algorithm's structure and the code and be able to easily modify the algorithm without disturbing other parts.

The code in its final form is provided as a **self-contained python class**, built for the deep-learning framework **pyTorch**. According to Python conventions, all the internal functions that a user should not call are marked as hidden by having names beginning with an "_". This is an implementation of

the design pattern *Low Coupling* and tells the user a clear interface. All functions have descriptive names. For example, in order to start the estimation process, the function `start_estimate()` needs to be called.

The class `Preconditioner` **inherits** from `torch.optim.Optimizer`. This means it follows all the conventions for how optimizers are expected to behave in pytorch. This makes it intuitive to use for the pytorch user. Features that were added to support this include the `state` dict, which can be used to save and load the state of the optimizer in order to pause or continue training. Another supported feature are parameter groups. This allows to treat different parameters seperately, for example different layers of a neural network. Specifically, each parameter group will be optimized with a seperate learning rate.

The class `Preconditioner` wraps an **inner optimizer** which is responsible for the actual parameter updates. In the original paper and in this thesis, only SGD is studied as an option, but this modification allows to use preconditioning and the adaptive learning rate estimation together with other optimizers which are implemented in a pytorch optimizer class. The Preconditioner class is responsible for estimating the Hessian. Once the estimate is complete, the class turns into a *decorator* for the provided inner optimizer class. Before the inner optimizer does its optimization step, the previously constructed preconditioner is applied to the parameters' gradient.

**Logging** data during training is a global task and should not be a responsibility of the optimizer. In order to expose the data of interest, the class exhibits a method `get_log()`, which returns some values of interest. It can be overwritten if the user wants to see other data. The user then takes care of further processing and writing the data to a file.

In order to change behavior of the class during development and research, the best way to make different versions is to make use of python's built-in **subclassing**. For example, in the experiments there is a version `AdaptiveSGD` that skips applying the preconditioner every step, but behaves exactly the same as `Preconditioner` in all other ways. `AdaptiveSGD` is a subclass of `Preconditioner` that overwrites the function `_apply_preconditioner()` with an empty function.

Some minor bugs were also fixed. Previously, the algorithm skipped some minibatches during the estimation phase. However, there remains room for improvement. The class assumes the user to input sensible values and won't complain otherwise. For some variables, helpful unit checks are in place. This is fine in a research environment, but not for deployment.

## 3.3 Test Problems

The optimizers were tested on some of the testproblems that come with DeepOBS and had a stable pytorch implementation at the time of testing. DeepOBS specifies the data set, the model architecture and initialization, the loss function and regularization. It also proposes a standard batch size that is shared by all baselines.

Before training, the data set is split into a validation set for hyperparameter tuning, a training set and a holdout test set.

The DeepOBS testproblem `fmnist_2c2d` uses the Fashion-MNIST data set ([Xiao et al., 2017]), which consists of 60000 labeled greyscale $32 \times 32$ pixel images of ten classes of fashion items. The model used consists of two convolutional layers, each followed by max-pooling, a fully connected layer with ReLU activation functions, and a 10-unit softmax output layer. The initial weights are drawn from a truncated normal distribution with $\sigma = 0.05$ and initial biases are set to 0.05. As a loss function, cross entropy loss is used, without any regularization term. The standard batch size is 128.

The DeepOBS testproblem `cifar10_3c3d` uses the CIFAR10 data set ([Krizhevsky et al., 2014]), which consists of 60000 labeled RGB $32 \times 32$ pixel images of ten classes of objects. The model used consists of three convolutional layers with ReLU activation functions, each followed by max-pooling, two fully connected layers with respectively 512 and 256 units and ReLU activation functions, and a 10-unit softmax output layer. The initial weights are initialized using Xavier initialization and initial biases are set to 0.0. As a loss function, cross entropy loss is used, with L2 regularization on the weights (but not the biases) with a factor of 0.002. The standard batch size is 128.

For the experiment investigating computational complexity, DeepOBS uses the built-in testproblem `mnist_mlp`, which is a multi-layer perceptron that consists of four layers with 1000, 500, 100 and 10 unis respectively, using ReLU activation on the first three layers and softmax on the output layer. Initial weights are drawn from a truncated normal distribution ($\sigma = 0.03$)and the initial biases are set to 0.0. As a loss function, cross entropy loss is used, without any regularization term. The standard batch size is 128.

In the experiment investigating different initialization methods, the goal was to replicate the findings from [Roos and Hennig, 2019], so the exact same model was used. It is very similar to the one that comes with DeepOBS. There are three convolutional layers, each followed by ReLU and maxpooling. After that, there are three fully-connected layers preceded by ReLUs. As a loss function, cross-entropy loss is used, with regularization as in `cifar10_3c3d`. Batch size and initialization methods are varied.

## 3.4 DeepOBS baselines

When using a standardized testing procedure like the one provided by DeepOBS, it is not necessary to run every optimizer, as reproducibility of the results is guaranteed across different hardware. This saves a lot of work and energy that every single researcher otherwise would have needed to spend. DeepOBS comes with results from many established and well-studied optimizers as a baseline, and it is easy to compare a new optimizer against them. Their hyperparameters are tuned for each testproblem, but some more complicated widely used techniques like hyperparameter schedules are not available.

## 3.5 Technical details

The experiments were run on the TCML cluster at the University of Tübingen. A Singularity container was set up on Ubuntu 16.4 LTS with python 3.5, pytorch (version) and DeepOBS (see Appendix for Singulariy recipe). Computation was distributed over multiple GPU compute nodes using the workload manager Slurm. Every job had access to 4 CPU cores (Intel XEON CPU E5-2650 v4), 3GB of RAM, and 1 GPU (GeForce GTX 1080 Ti).

During development, I used git on github for distributed version control.

# Chapter 4

# Experiments

The main goal of the experiments was to investigate the training performance of the proposed algorithm on neural networks. In Experiments 1 and 2, the various phases of the optimizer were tested seperately for training performance and wallclock time as a proxy for complexity. In Experiments 3 and 4, outside factors were varied to study the influence of initialization, batch size and the first-epoch learning rate on the proposed full algorithm.

Experiments were run on the setup described in the previous chapter, using DeepOBS standard procedures and parameters.

## 4.1 Experiment 1: Preconditioning

This experiment consists of two parts: In part A, the full preconditioning algorithm applied to SGD is tested against the baselines. In part B, the different phases of the algorithm are tested seperately by using subclasses of `Preconditioner` that knock out the corresponding functions. The hyperparameters used during the runs were the same for all versions of the algorithm: `num_observations = 10`, `prior_iterations = 5`, `est_rank = 2`, `optim_class = torch.SGD`, `lr = None`. The DeepOBS test problems used were `fmnist_2c2d` and `cifar10_3c3d`.

The results of part A are presented in figure 4.1, together with the DeepOBS baselines on the same problems. In almost every metric, Preconditioned SGD is outperformed by all widely used baselines like bare SGD, Momentum and Adam. Interestingly, the `fmnist_2c2d` problem seems prone to overfitting: All the standard optimizers increase on test loss after reaching a local minimum. However, the test accuracy does not seem to suffer from this same problem. Preconditioned SGD does not show this behavior, but also does not reach convergence even after 100 epochs, when they still continue to improve, while the baseline optimizers reach convergence between 40 an d 60 epochs. Another

finding is that the variance of Preconditioned SGD is larger than the variance of the baseline optimizers.

For part B of the experiment, different versions of the preconditioning algorithm were created. See the appendix for an example definition of `AdaptiveSGD`. Table 4.1 shows the definitions of the variations. The "Tuned" versions use a constant learning rate during the entire run, precisely the one which was tuned for the baseline standalone inner optimizer for each test problem. Momentum was run with the parameter `momentum = 0.9`, as in the DeepOBS baseline.

| | (1) | (2) | (3) | (4) | Inner optimizer |
|---|---|---|---|---|---|
| Preconditioned SGD | ✓ | ✓ | ✓ | ✓ | SGD |
| Adaptive SGD | ✓ | ✓ | ✓ | | SGD |
| Preconditioned Tuned Momentum | ✓ | | ✓ | ✓ | Momentum |
| Preconditoned Tuned SGD | ✓ | | ✓ | ✓ | SGD |

**Table 4.1:** The original algorithm, Preconditioned SGD, (1) calculates a prior, (2) constructs and uses a learning rate for the inner optimizer, (3) calculates the preconditioner and (4) applies it before every step of the inner optimizer. The variations are defined as shown here.

Figure 4.2 shows the results of part B. On every metric, Preconditioned SGD is outperformed by the variant without preconditioning, Adaptive SGD, but both perform much worse than the other optimizers. On cifar10_3c3d, the "Tuned" versions perform very similarly to standalone Momentum, beating standalone SGD. On fmnist_2c2d, the "Tuned" versions perform similarly to their respective standalone optimizers, but slightly worse. The overfitting behaviour on fmnist_2c2d observed in part A is also exhibited by the "Tuned" optimizers.

## 4.2 Experiment 2: Wallclock Time

To the practicioner it is important how much time, energy and money the training of a model costs. For this reason, DeepOBS includes a script to estimate wallclock time for given optimizers as compared to SGD. Wallclock time is the runtime as measured by a clock on the wall. It does not correct for hardware and parallelization. As such, it is not a direct measurement for an algorithm's complexity, but rather a proxy for the efficiency of a specific implementation.

The three investigated optimizers were Preconditioned SGD and Adaptive SGD, as defined in table 4.1, and OnlyAdaptiveSGD, which is an optimized version of AdaptiveSGD. It simply skips the construction of the preconditioner, which is not applied anyways. As a further baseline, Adam's runtime is tested.

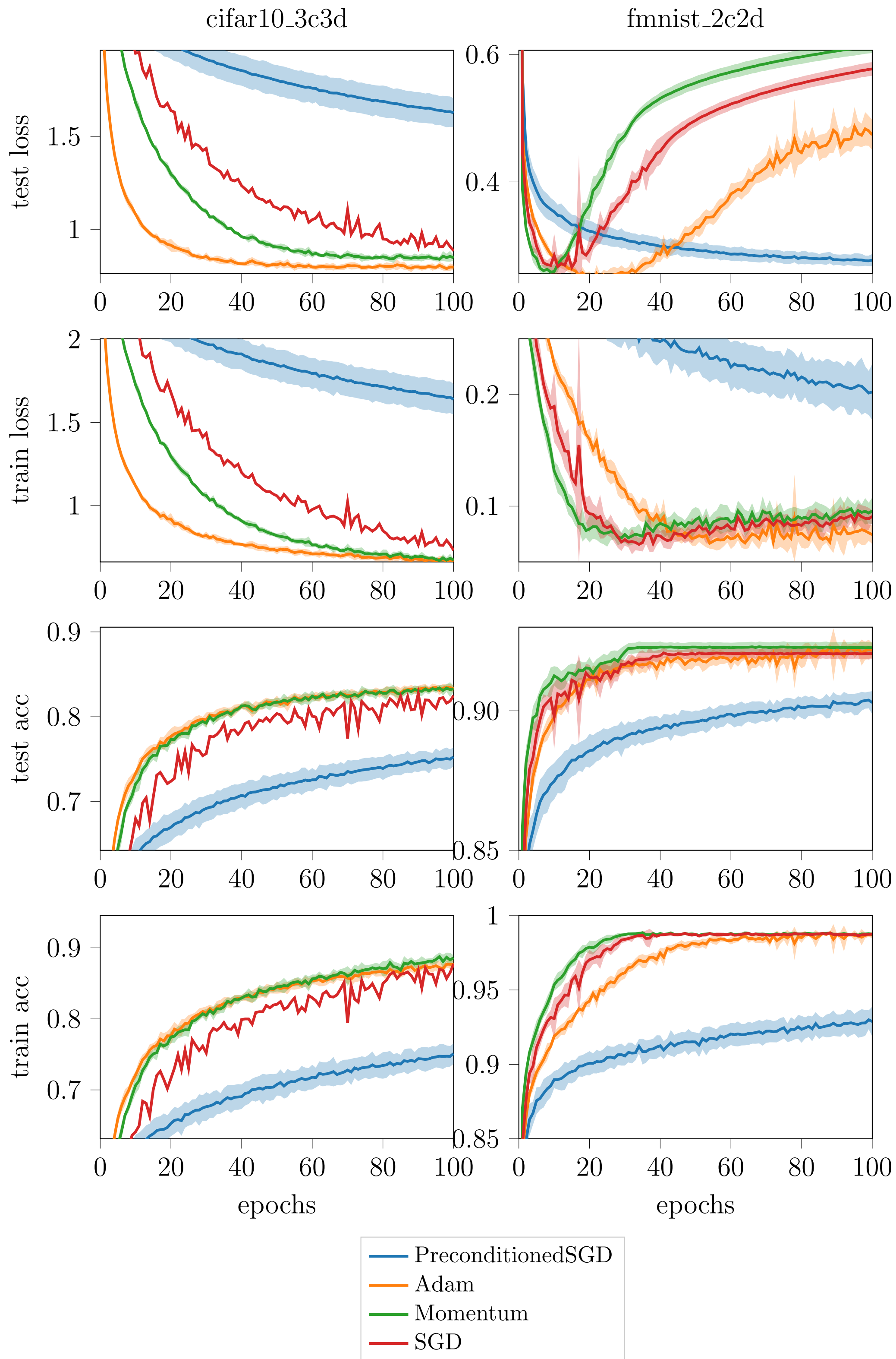


**Figure 4.1:** This figure shows training performance of Preconditioned SGD compared with DeepOBS baseline optimizers. It performs comparatively bad on both testproblems.

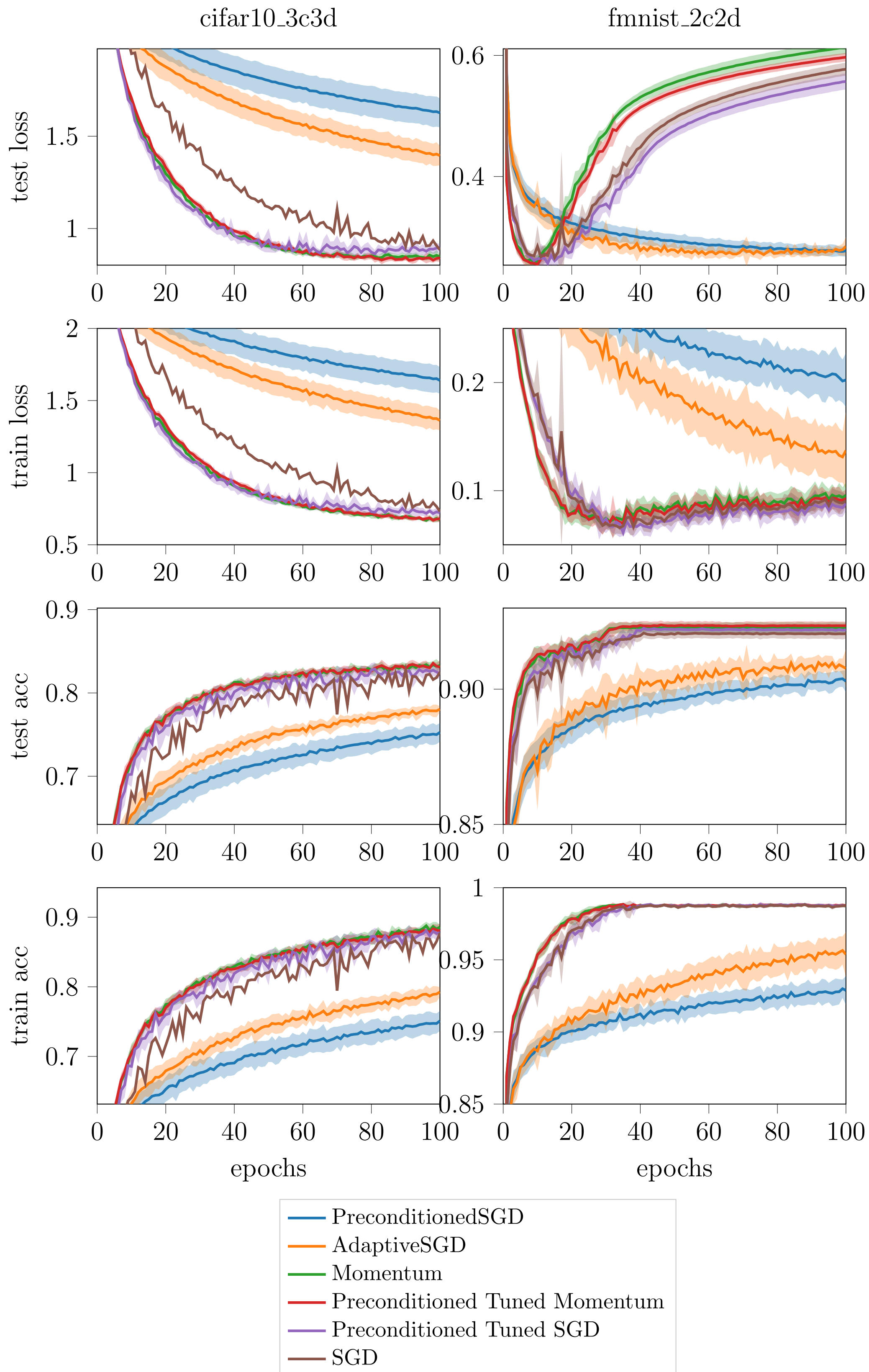


**Figure 4.2:** Variations of the original algorithm perform better. "Tuned" versions use the same learning rate as tuned baseline standalone optimizers. This figure shows the preconditioning algorithm, modified to use the same learning rate as well-tuned SGD, for every epoch. It performs better than standard tuned SGD using the same learning rate, and even better than the adaptive versions.

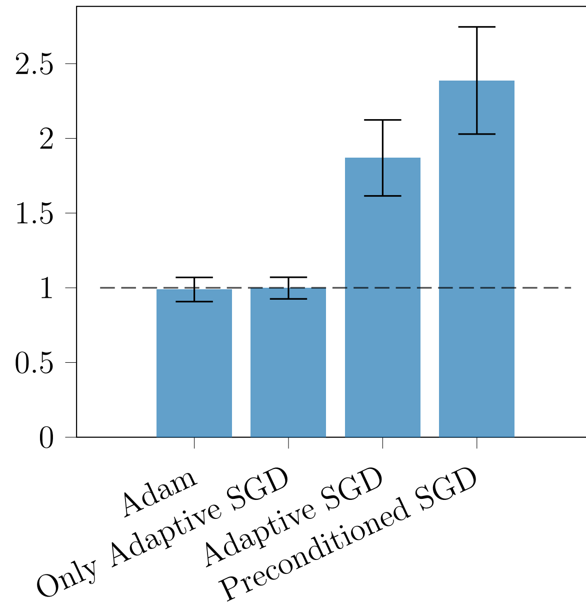


**Figure 4.3:** Experiment 2: Wallclock time per epoch, as tested on the DeepOBS testproblem mnist_mlp. A score of 1 represents the baseline runtime of SGD. The original algorithm is the slowest. Calculating, but not applying the preconditioner is considerably faster. Only gathering the prior observations is about as fast as SGD and Adam.

The testproblem used was the default value set by DeepOBS, `mnist_mlp`, which is a rather small model on mnist, using a batch size of 128. Every optimizer was run 5 times for 5 epochs each, on a GPU node of the TCML cluster.

The results are presented in figure 4.3. As expected, the more phases of the algorithm the preconditioner version goes through, the more time it takes to run. `PreconditionedSGD` was the slowest optimizer, taking an mean of 2.39 ($SD = 0.36$) times as long as SGD, followed by `AdaptiveSGD`, which required 1.87 ($SD = 0.25$) times the runtime of SGD. `OnlyAdaptiveSGD` took about as much time as SGD, with a mean time of 1.00 ($SD = 0.07$) the time of SGD. `Adam` took about as long as SGD, with a mean time of 0.99 ($SD = 0.08$).

## 4.3 Experiment 3: Initialization

The original algorithm as studied in [Roos and Hennig, 2019] includes a hyperparameter for the manual first-epoch learning rate, because the learning rate constructed by the algorithm would be so high that the model becomes unstable and diverges in the first epochs. Using the implementation proposed in this thesis, this instability was not observed. On all tested problems included

in DeepOBS, using the constructed learning rate for the first epoch lead the algorithm to eventual convergence.

The goal for the experiment was to narrow down the source of the instability by replicating the original setup as closely as possible. The DeepOBS testproblem `cifar10_3c3d` differs in three ways from the original setup. DeepOBS uses an explicit initialization of the weights, while the original setup uses the pyTorch default methods. DeepOBS sets the default batch size for the cifar10 problems to 128, while the original paper used 32. The model itself also differs in small ways.

For this experiment, DeepOBS was adapted to use the original model and the initialization method was varied, as well as the batch size. In order to be able to spot inconsistency, each run was 5 epochs long and was repeated on 10 different random seeds. The used optimizer version was Preconditioned SGD, as in the experiments before.

The results are presented in figure 4.4. Both the initialization method and batch size have an effect on training performance, but the algorithm was stable for all combinations. The pyTorch initialization is a better prior for this test problem than the DeepOBS initialization. The smaller the batch size, the better Preconditioned SGD performs.

## 4.4 Experiment 4: Learning Rate Sensitivity

The proposed algorithm constructs a learning rate for the inner optimizer, but one can manually specify a learning rate for the first epoch. The goal of this experiment was to test how the choice of first-epoch learning rate influences training performance of Preconditioned SGD.

DeepOBS was used to create independent commands for a grid search with 10 evaluations on a logarithmic grid between $10^-5$ and $10^2$ on the `fmnist_2c2d` testproblem. Having independent commands to execute on the computing cluster made the grid search easily parallelizable.

The results are shown in 4.5, together with the baseline SGD as reference and a dashed line which represents Preconditioned SGD constructing the first-epoch learning rate by itself. Like SGD, the model diverges in the first epoch if the learning rate is set over a certain threshold. Below this threshold, the first-epoch learning rate does not seem to have an effect on the model's final accuracy after training for 100 epochs. For SGD, which uses the same learning rate for all 100 epochs, there is a significant dropoff in training success for smaller learning rates. Both SGD and Preconditioned SGD with a stable manual first-epoch learning rate achieve a similar final accuracy to Preconditioned SGD using the automatically constructed learning rate.

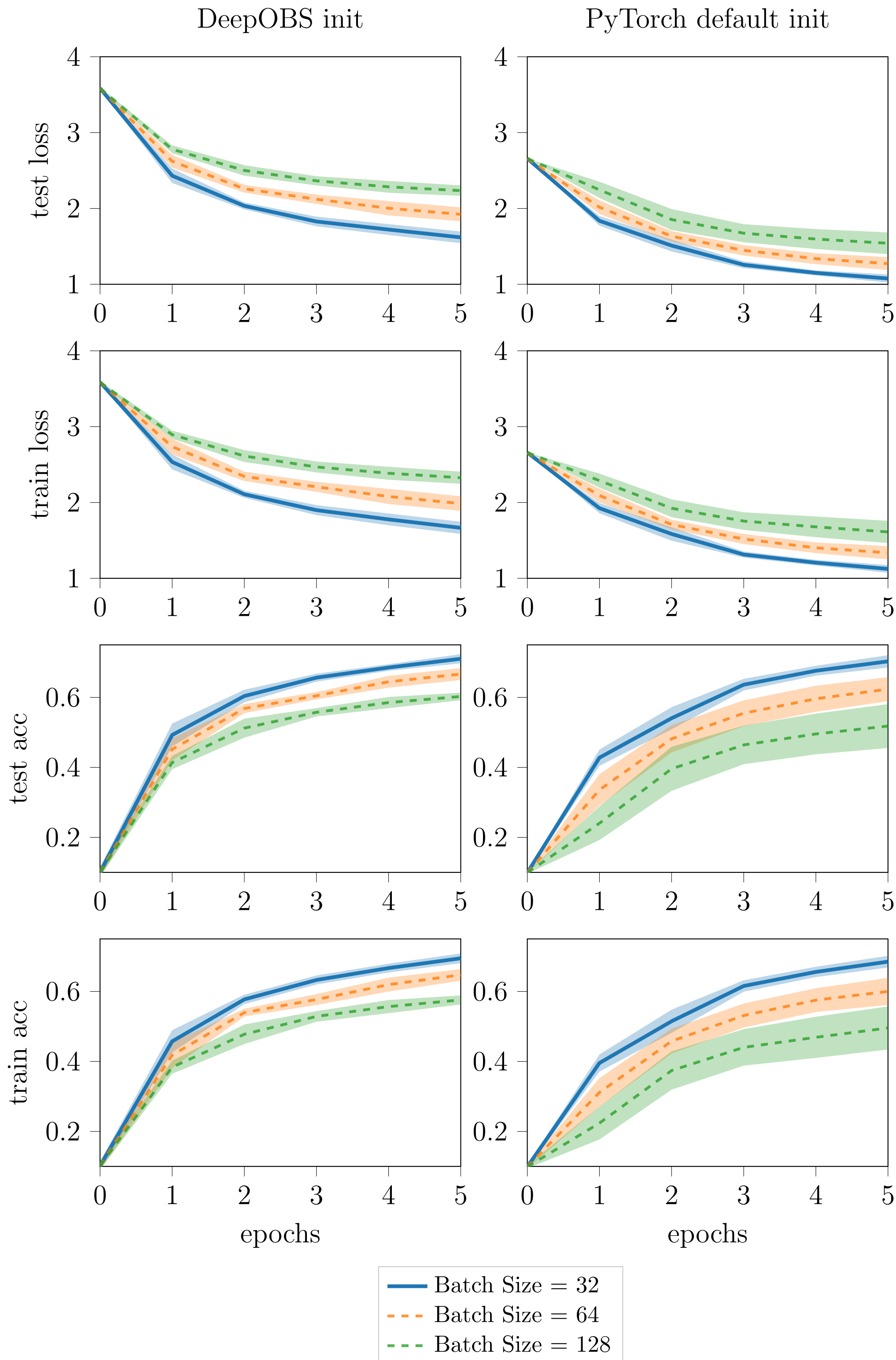


**Figure 4.4:** Experiment 3: PreconditionedSGD on a cifar10 net, comparing initialization methods. The algorithm is stable regardless of batch size and initialization method.

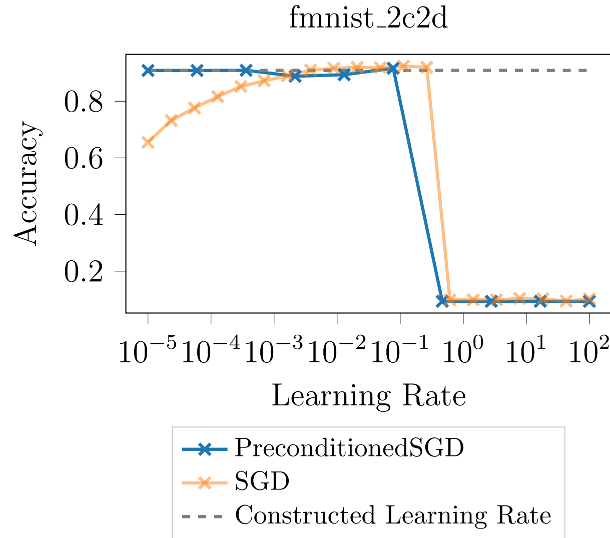


**Figure 4.5:** Experiment 4: The originally proposed algorithm allows the user to specify the inner optimizer's (in this case SGD) learning rate in the first epoch. Setting it below a certain model-specific threshold, the algorithm performs just as well as the constructed learning rate. Setting it over that threshold, the algorithm diverges in the first epoch.

## 4.5 Discussion

The experiments presented in the previous sections show that the tested implementation of the algorithm proposed in [Roos and Hennig, 2019] is not generally useful for application in deep learning.

In Experiment 1A the full algorithm performs worse than established baseline optimizers on nearly every measure of training success. Experiment 1B was set up to isolate the individual parts of the algorithm and investigate them seperately.

The fmnist test problem seems prone to overfitting. After reaching a very good point, test loss of the baseline optimizers increased, while train loss remained low and accuracies high. The preconditioning algorithm does not show this behaviour. This can be explained with the generally slow training speed. After 100 epochs, it doesn't even reach the minimum that the baselines have achieved very quickly.

During its first phase, the algorithm gathers information about the curvature of the loss landscape and uses this information to construct a diagonal prior estimate for the Hessian. This is then used to construct a learning rate for the inner optimizer. The step size changes every epoch and thereby allows the optimizer to respond to changes in local curvature. However, using a fixed tuned learning rate outperforms the adaptive learning rate in every case.

Basing the constructed learning rate on only the first minibatches per epoch makes the optimizer less robust. Training success of the whole epoch depends largely on the learning rate, which is heavily influenced by sampling noise. This explains the increased variance of the adaptive optimizers in figure 4.2. In experiment 3, the batch size is varied. A smaller batch size increases sampling noise, but surprisingly, it decreases variance of the optimizer's performance. The benefit of taking more steps seems to outperform increased sampling noise on robustness.

The adaptive learning rate is not robust and it performs worse than a tuned fixed learning rate, but might still be more effective than going through an extensive tuning process.

The algorithm applies the preconditioner before every optimization step performed by the inner optimizer. Applying the preconditioner to SGD while using a fixed tuned learning rate lead to an increase in performance one one of the two test problems. When using Tuned Momentum or when running on the other test problem, applying the preconditioner had a very small effect, if any.

When using the constructed learning rate on SGD, applying the preconditioner lead to a decreased performance.

The preconditioning algorithm did not perform better than well-tuned momentum, which seems able to better capture local curvature with much less computational effort.

There are limitations to this study that might contribute to the preconditioning algorithm underperforming compared to the original paper. In the original paper, the used testing protocol is not thoroughly explained. Test problem, batch size and SGD learning rates were set manually and without justification, which makes the results hard to interpret. In a similar way, the Preconditioner's hyperparameters were not tuned in a systematic way. Trying to replicate the paper's result as closely as possible in Experiment 3, I failed to reproduce the observed instability. This points to a functional difference between implementations.

Given the theory, Preconditioned Tuned SGD should reliably outperform standalone Tuned SGD. It however has a noticeable effect only on the cifar10 testproblem. There are multiple factors that might be contributing to this.

The main factor is that on the fmnist problem, standalone SGD performs comparatively to Momentum. This means that the problem is already quite well-conditioned and the scalar learning rate is able to capture global curvature.

Another possible reason that SGD is not outperformed by the preconditioner is that while estimating the Hessian, the Preconditioner does not use the data for actual parameter updates. The effect of this depends on the number of minibatches.

In experiment 2, the time penalty of preconditioning was investigated. The

authors report that the time of building the rank 2 approximation accounted for 2-5% of the total computational cost per epoch, using a batch size of 32 on the training subset of CIFAR-10. This number does not tell the whole story. This thesis adds to this finding the performance penalty of estimating the Hessian and applying the preconditioner. Gathering the prior observations add no considerable overhead compared to SGD, while both calculating and applying the Preconditioner add considerable overhead. Adam as another adaptive method is almost exactly as fast as SGD.

Using the DeepOBS standard batch size of 128, there are fewer minibatches per epoch, which means the estimation of the Hessian on the same number of steps uses a larger proportion of total training time. In this setting, using the preconditioner incurs a computational cost of more than double that of SGD. Two epochs of SGD can be run in the time it takes to compute the preconditioner and apply it during every step in one epoch.

The implementation of the algorithm also plays a large role. As construction of the preconditioner is not parallelized on the GPU, it is expected to take a comparatively long time. However, applying the preconditioner every step is a GPU operation and increases epoch time by about 50% of an SGD epoch.

Measuring CPU time would also be interesting, because that would account for the algorithm not being fully GPU-optimized. Unfortunately, DeepOBS does not provide this functionality.

Experiment 3 tried to replicate the author's findings that the algorithm would construct a very large learning rate in the first epoch, causing it to fail. Using the same model, batch size and initialization method as in the original paper, the new implementation proved stable. Interestingly, the algorithm produces greater variance on accuracies when using larger batch sizes.

The runs in experiment 3 were only done for a small number of test problem/hyperparameter combinations. There is no guarantee that the found stability translates to other kinds of problems, so it is useful to keep the option of setting an initial learning rate. It is also unclear exactly why the new algorithm is stable.

Experiment 4 investigated the manual first-epoch learning rate. If set above a certain threshold, the algorithm diverges in the first epoch. If set below this threshold, the algorithm does not diverge. After 100 epochs, there is little to no effect of the first-epoch learning rate. SGD is more sensitive to the learning rate, where a low choice lead to considerably worse training performance.

If the algorithm is used to construct a learning rate, one should therefore not need to specify a first-epoch learning rate. However, given the poor performance of the constructed learning rate observed in experiment 1, it might be a good idea to specify a tuned fixed learning rate throughout all epochs.

Generally, the experiments in this thesis were mostly performed on convolutional nets, and it is unclear which of the found properties would translate

to other architectures.

The usability goals of the implementation were met, as the experiments required no modifications to the final optimizer class to be run and integrate nicely with DeepOBS.

## 4.6 Further Research and Development

This thesis can not exhaustively cover all aspects of the presented optimizer implementation. There are open questions the experiments did not answer and several improvements to be made to the implementation and the algorithm itself.

The relationship between the Preconditioner's remaining hyperparameters and accuracy and computational overhead remains unclear. How much better does the estimate get when using an additional data point? The Hessian is taken as the mean of a Gaussian distribution, so the variance of this distribution could be taken as a measure for uncertainty. The algorithm could keep using new data points until the certainty of the estimate does not improve anymore and the remaining variance can mostly be explained by minibatch sampling noise.

A related open question is when to best restart the estimation process. This could be at the start of every epoch, as throughout this thesis, after some fixed number of minibatches or dynamically, once the estimate is determined to be a bad fit at the current point in parameter space. Estimating the likelihood of the Hessian given the observed gradients could also be achieved using the variance of the Hessian estimate, which isn not calculated in the current version.

It is also unclear whether parameter groups are only a convenient feature for practicioners or whether using them does actually impact training success.

In order to make general statements about the robustness and performance of the algorithm, it would need to be either further examined analytically or run on many more, different kinds of model architectures.

## 4.7 Feedback for DeepOBS

This thesis is one of the first semi-external projects that use DeepOBS and more specifically the pyTorch version of DeepOBS. Naturally, some issues and unexpected behavior came up while using it. This is a selection of Feedback on DeepOBS.

- Usability:
  - In general, DeepOBS is quite easy to use. The minimal examples provided with the documentation work well and are easy to adapt

to other optimizers.
  - Output files used to be single-line strings of json. This was changed to easily human-readable formatting.
  - The way DeepOBS currently handles output files is very transparent (It simply writes json files for each optimizer run in an appropriate file structure). However, there is no unique place where DeepOBS saves and looks for metadata. Sometimes it is in the folder name, sometimes in the file itself, where it should be. It would make more sense to save consistently defined "run"-objects which point to all the runs that logically belong together.
  - When analyzing runs, DeepOBS should not rely on the operating system to determine things like the order of testproblems. The user should be able to define this, for example by manipulating a "run" object before plotting.
- Protocol:
  - The testproblems' hyperparameters like the batch size do affect optimizer performance a lot. DeepOBS provides default values, but it should be clearer if and how they can be left on their default values to compare optimizers without introducing bias.
  - DeepOBS now uses a web-based issue tracking system where users can easily report bugs and request additional functionality.
- Features:
  - DeepOBS lacks a testproblem that uses parameter groups. While this might be less important for optimizer testing, it would be useful for optimizer development.
  - For measuring computational complexity, CPU time might be a more useful metric than Wallclock time, because that would rely less on the algorithm being tuned to the used hardware setup.
  - In the output files, DeepOBS should record if a run was stopped (and why) or if the algorithm was stable all the way throughout the run.

# Chapter 5

# Conclusion

In this thesis, I present an implementation of the probabilistic preconditioning algorithm proposed in [Roos and Hennig, 2019]. This implementation is a python class built for pyTorch, which wraps an existing optimizer. It is straightforward to use and can be swapped out from another optimizer with a few lines of code.

The algorithm gathers observations of local curvature to estimate the Hessian of the Loss function. A low-rank estimate of this Hessian is then inverted and used as a preconditioner. Also, the algorithm constructs a learning rate for the inner optimizer from the observed curvature information. Once the preconditioner is constructed, it is applied to the observed gradient before each optimizer step.

Using the optimizer benchmarking framework DeepOBS, I show that Preconditioned SGD performs worse than well-tuned standard optimizers like SGD and Adam on convolutional neural networks. This effect is largely caused by the non-robust adaptive learning rate. Using a tuned fixed learning rate through all epochs for the inner optimizer lead to better performances than standalone SGD, though not exceeding the performance of Momentum.

As DeepOBS is actively being worked on, I provide some feedback. While the core functionality is quite mature, some areas like the handling of output files and the automatic visualizer leave room for further improvement.

Future implementations of the tested preconditioning algorithm should be fully GPU-optimized, which this one is not. If the algorithm can be further improved, especially by finding a more robust alternative for the adaptive learning rate, it can be a promising alternative to established optimizers.

# Appendix A

# Code

## A.1 Singularity build recipe

```
#header
Bootstrap: docker
From: ubuntu:16.04

#Sections

%environment
# set environment variables

%post
# commands executed inside the container after os has been installed.
# Used for setup of the container

apt-get -y update
apt-get -y install python3-pip git python3-tk

python3 --version

pip3 install --upgrade pip
pip install torch torchvision
pip install git+https://github.com/abahde/DeepOBS.git@v2.0.0-beta#egg=deepobs
```

This build recipe was used to build the Singularity container by invoking `sudo singularity build`

## A.2 DeepOBS Runscript for experiment 1

This code includes the definitions for PreconditionedSGD and AdaptiveSGD. It was run 10 times using different seeds each time.

```
"""Simple run script using SORunner."""

import torch.optim as optim
import deepobs.pytorch as pyt
from sorunner import SORunner
from probprec import Preconditioner
import numpy
import math

# DeepOBS setup

class PreconditionedSGD(Preconditioner):
    """docstring for PreconditionedSGD"""
    def __init__(self, *args, **kwargs):
    super(PreconditionedSGD, self).__init__(
        *args,
        optim_class = optim.SGD,
        **kwargs)

class AdaptiveSGD(Preconditioner):
    """docstring for PreconditionedSGD"""
    def __init__(self, *args, **kwargs):
    super(AdaptiveSGD, self).__init__(
        *args,
        optim_class = optim.SGD,
        **kwargs)

    def _apply_preconditioner(self):
        return;


# specify the Preconditioned Optimizer class
poptimizer_class = PreconditionedSGD

# and its hyperparameters
phyperparams = {'lr': {"type": float, 'default': None}}

# create the runner instances
```

```
prunner = SORunner(PreconditionedSGD, phyperparams)

runner = SORunner(AdaptiveSGD, phyperparams)

runner.run(testproblem='fmnist_2c2d')
runner.run(testproblem='cifar10_3c3d')
prunner.run(testproblem='fmnist_2c2d')
prunner.run(testproblem='cifar10_3c3d')
```

## A.3 Slurm Batch Definition File

This file was used to run experiment 1 on the compute cluster.

```
#!/bin/bash

####
#a) Define slurm job parameters
####

#SBATCH --job-name=prec

#resources:

#SBATCH --cpus-per-task=4
# the job can use and see 4 CPUs (from max 24).

#SBATCH --partition=day
# the slurm partition the job is queued to.

#SBATCH --mem-per-cpu=3G
# the job will need 12GB of memory equally distributed on 4 cpus.
# (251GB are available in total on one node)

#SBATCH --gres=gpu:1
#the job can use and see 1 GPUs (4 GPUs are available in total on one node)

#SBATCH --time=03:30:00
# the maximum time the scripts needs to run
# "hours:minutes:seconds"

#SBATCH --array=43-51
```

```
#SBATCH --error=job.%J.err
# write the error output to job.*jobID*.err

#SBATCH --output=job.%J.out
# write the standard output to job.*jobID*.out

#SBATCH --mail-type=ALL
#write a mail if a job begins, ends, fails, gets requeued or stages out

#SBATCH --mail-user=ludwig.bald@student.uni-tuebingen.de
# your mail address

####
#b) copy all needed data to the jobs scratch folder
####

DATA_DIR=/scratch/$SLURM_JOB_ID/data_deepobs/

mkdir -p $DATA_DIR/pytorch
mkdir -p $DATA_DIR/pytorch/FashionMNIST/raw
cp -R /common/datasets/cifar10/. $DATA_DIR/pytorch/
cp -R /common/datasets/Fashion-MNIST/. $DATA_DIR/pytorch/FashionMNIST/raw/

####
#c) Execute the code
####


singularity exec ~/image.sif python3 ~/exp_preconditioning/runscript.py
    --data_dir $DATA_DIR --random_seed $SLURM_ARRAY_TASK_ID

echo DONE!
```

## A.4 Repository

Everything that's needed to reproduce the experiments in this thesis is freely accessible. The python and bash source code for the preconditioner, cluster use, experiment setup are available in the following repository: `https://github.com/ludwigbald/probprec/` Raw experiment result files and the LaTeXsources for this thesis and the presentation are in the same repository.

# Bibliography

[Bahde, 2019] Bahde, A. (2019). Tuning procedure for deepobs. Master's thesis, Universität Tübingen. Master thesis.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

[Brown and Hamarneh, 2016] Brown, C. J. and Hamarneh, G. (2016). Machine learning on human connectome data from mri. *arXiv preprint arXiv:1611.08699*.

[Chaudhari et al., 2017] Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J. T., Sagun, L., and Zecchina, R. (2017). Entropy-sgd: Biasing gradient descent into wide valleys. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

[Gibney, 2016] Gibney, E. (2016). Google ai algorithm masters ancient game of go. *Nature News*, 529(7587):445.

[Krizhevsky et al., 2014] Krizhevsky, A., Nair, V., and Hinton, G. (2014). The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55.

[Roos and Hennig, 2019] Roos, F. and Hennig, P. (2019). Active probabilistic inference on matrices for pre-conditioning in stochastic optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1448–1457.

[Saad, 2003] Saad, Y. (2003). *Iterative methods for sparse linear systems*, volume 82. siam.

[Schneider et al., 2019] Schneider, F., Balles, L., and Hennig, P. (2019). Deepobs: A deep learning optimizer benchmark suite. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

[Vinyals et al., 2019] Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W. M., Dudzik, A., Huang, A., Georgiev, P., Powell, R., et al. (2019). Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind Blog*.

[Xiao et al., 2017] Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

# Selbstständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

---

Tübingen, 14. Oktober 2019 Ludwig Bald