

ADAHESIAN: An Adaptive Second Order Optimizer for Machine Learning

Zhewei Yao^{*,1}, Amir Gholami^{*,1}, Sheng Shen¹, Mustafa Mustafa², Kurt Keutzer¹, Michael W. Mahoney¹

¹University of California, Berkeley, ²NERSC, Lawrence Berkeley National Laboratory
{zhewei, amirgh, sheng.s, keutzer, mahoneymw}@berkeley.edu, mmustafa@lbl.gov

Abstract—We introduce ADAHESIAN, a second order stochastic optimization algorithm which dynamically incorporates the curvature of the loss function via ADaptive estimates of the HESSIAN. Second order algorithms are among the most powerful optimization algorithms with superior convergence properties as compared to first order methods such as SGD and Adam. The main disadvantage of traditional second order methods is their heavier per-iteration computation and poor accuracy as compared to first order methods. To address these, we incorporate several novel approaches in ADAHESIAN, including: (i) a fast Hutchinson based method to approximate the curvature matrix with low computational overhead; (ii) a root-mean-square exponential moving average to smooth out variations of the Hessian diagonal across different iterations; and (iii) a block diagonal averaging to reduce the variance of Hessian diagonal elements. We show that ADAHESIAN achieves new state-of-the-art results by a large margin as compared to other adaptive optimization methods, including variants of Adam. In particular, we perform extensive tests on CV, NLP, and recommendation system tasks and find that ADAHESIAN: (i) achieves 1.80%/1.45% higher accuracy on ResNets20/32 on Cifar10, and 5.55% higher accuracy on ImageNet as compared to Adam; (ii) outperforms AdamW for transformers by 0.13/0.33 BLEU score on IWSLT14/WMT14 and 2.7/1.0 PPL on PTB/Wikitext-103; (iii) outperforms AdamW for SqueezeBert by 0.41 points on GLUE; and (iv) achieves 0.032% better score than Adagrad for DLRM on the Criteo Ad Kaggle dataset. Importantly, we show that the cost per iteration of ADAHESIAN is comparable to first-order methods, and that it exhibits robustness towards its hyperparameters. The code for ADAHESIAN is open-sourced and publicly-available [1].

I. INTRODUCTION

The high dimensional and non-convex nature of many machine learning tasks has rendered many classical optimization methods inefficient for training and/or evaluating Neural Network (NN) models. After decades of

research, first order methods, and in particular variants of Stochastic Gradient Descent (SGD), have become the main workhorse for training NN models. However, they are by no means an ideal solution for training NN models. There are often a lot of ad-hoc rules that need to be followed very precisely to converge (hopefully) to a point with good generalization properties. Even the choice of the first order optimizer has become an ad-hoc rule which can significantly affect the performance. For example, SGD with momentum is typically used in Computer Vision (CV); Adam is used for training transformer models for Natural Language Processing (NLP); and Adagrad is used for Recommendation Systems (RecSys). Using the wrong SGD variant can lead to significant performance degradation. Another challenging ad-hoc rule is the choice of hyperparameters and hyperparameter tuning methods, even after an optimizer is chosen. Hyperparameters include learning rate, decay schedule, choice of momentum parameters, number of warmup iterations, etc. As a result of these and other issues, one has to *babysit* the optimizer to make sure that training converges to an *acceptable* training loss, without any guarantee that a given number of iterations is enough to reach a local minima.

Importantly, one may *not* observe the above problems for certain popular learning tasks, such as ResNet50 training on ImageNet. The reason is that, for these tasks, years of industrial scale hyperparameter tuning has lead to what may be called *ideal SGD behaviour*. That is, for this problem, hyperparameters have been brute-force engineered to compensate for the deficiencies of first order methods. Such a brute force approach is computationally and financially not possible for many large scale learning problems—certainly it is not possible to do routinely—and this has made it challenging to train and apply NN models reliably.

Many of these issues stem from the fact that first order methods only use gradient information and do not

*Equal contribution.

consider the curvature properties of the loss landscape, thereby leading to their suboptimal behaviour. Second order methods, on the other hand, are specifically designed to capture and exploit the curvature of the loss landscape and to incorporate both gradient and Hessian information. They are among the most powerful optimization algorithms, and they have many favorable properties such as resiliency to *ill-conditioned* loss landscapes, invariance to parameter scaling, and robustness to hyperparameter tuning. The main idea underlying second order methods involves *preconditioning* the gradient vector before using it for weight update. This has a very intuitive motivation related to the curvature of the loss landscape. For a general problem, different parameter dimensions exhibit different curvature properties. For example, the loss could be very flat in one dimension and very sharp in another. As a result, the step size taken by the optimizer should be different for these dimensions, and we would prefer to take bigger steps for the flatter directions and relatively smaller steps for the sharper directions. This can be illustrated with a simple 2D quadratic function as shown in Figure 1, where we show the trajectories of different optimizers. As can be seen, first order methods need a large number of steps for convergence and/or are hard to converge at all without hyperparameter tuning. However, second order methods capture this curvature difference, by normalizing different dimensions through rotation and scaling of the gradient vector before the weight update. Nonetheless, this comes at a cost. Despite the theoretically faster convergence rate of second order methods, they are rarely used for training NN models. This is due in part to their high computational cost.

In this paper, however, we will show that it is possible to compute *approximately* an exponential moving average of the Hessian and use it to precondition the gradient adaptively. The result is ADAHESSIAN, an adaptive optimizer that exceeds the state-of-the-art performance for a wide range of learning problems, including ResNets [25] for CV, transformers [43] for NLP problems, and DLRM [40] models for RecSys tasks. In more detail, the main contributions of our work are the following.

- 1) To reduce the overhead of second order methods, we approximate the Hessian as a diagonal operator. This is achieved by applying Hutchinson’s method to approximate the diagonal of the Hessian. Importantly, this approximation allows us to efficiently apply a root-mean-square exponential moving average to smooth out “rugged” loss surfaces. The advantage of this approach is that it has $\mathcal{O}(d)$ memory complexity.

- 2) We incorporate a block diagonal averaging to reduce the variance of Hessian diagonal elements. In particular, this has no additional computational overhead in the Hutchinson’s method, but it favorably affects the performance of the optimizer.
- 3) To reduce ADAHESSIAN overhead, we measure the sensitivity of ADAHESSIAN to different hyperparameters such as learning rate, block diagonal averaging size, and delayed Hessian computation. Interestingly, our results show that ADAHESSIAN is robust to those hyperparameters. See Section V-A and V-B.
- 4) We extensively test ADAHESSIAN on a wide range of learning tasks. In all tests, ADAHESSIAN significantly outperforms other adaptive optimization methods. Importantly note that these results are achieved even though we use the same learning rate schedule, weight decay, warmup schedule, dropout, as well as first/second order moment coefficients. In particular, we consider the following tasks.
 - a) **Computer Vision:** ADAHESSIAN achieves significantly higher accuracy, as compared to Adam. For instance, for ResNet32 on Cifar10, ADAHESSIAN achieves 93.08% as opposed to 91.63% achieved by Adam. Furthermore, for ResNet18 on ImageNet, ADAHESSIAN achieves 70.08% accuracy as opposed to 64.53% of Adam. For all tests, ADAHESSIAN achieves similar performance to the ideal SGD behavior, which is a result of hyperparameters having been tuned for many years at the industrial scale. Comparison with other optimizers and other models is discussed in Section IV-B.
 - b) **Natural Language Processing:** ADAHESSIAN improves the performance of transformers for machine translation and language modeling tasks, as compared to AdamW. In particular, ADAHESSIAN significantly outperforms AdamW by 0.13/0.33 BLEU on IWSLT14/WMT14, and by 2.7/1.0 PPL on PTB/WikiText-103. Moreover, for SqueezeBERT [26] fine-tuning on GLUE, ADAHESSIAN achieves 0.41 better points than AdamW. See Section IV-C, IV-D, and IV-E for more details.
 - c) **Recommendation System:** ADAHESSIAN improves the performance of DLRM on the Criteo Ad Kaggle dataset by 0.032% as compared to Adagrad, which is commonly used. See Section IV-F for more details.
- 5) We measure the sensitivity of ADAHESSIAN to different hyperparameters such as learning rate, spatial

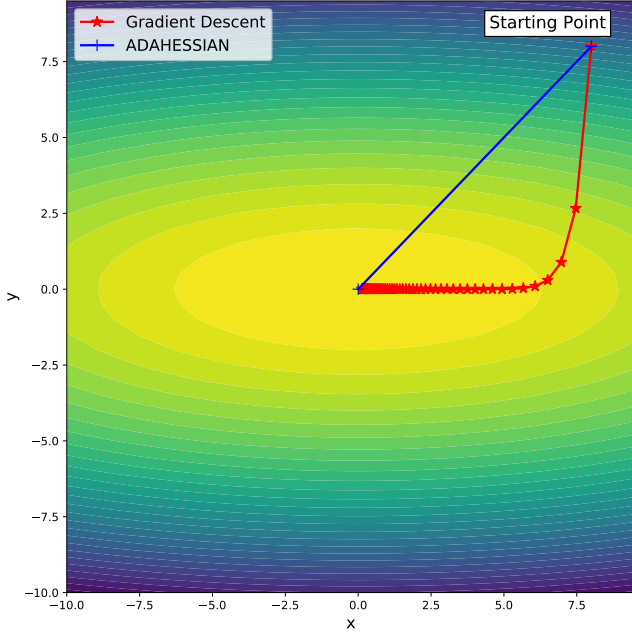


Fig. 1: The trajectory of gradient descent and ADAHESSIAN on a simple 2D quadratic function $f(x, y) = 10x^2 + y^2$. Gradient descent converges very slowly, even though this problem has a reasonable condition number. However, ADAHESSIAN converges to the optimum in just one step. This is because second order methods normalize the curvature difference between x and y axis by preconditioning the gradient vector before the weight update (by rescaling and rotating the gradient vector).

averaging size, and delayed Hessian computation. Interestingly, our results show that ADAHESSIAN is robust to those hyperparameters. See Section V-A and V-B for more details.

We emphasize that our empirical results are achieved even though we use the same learning rate schedule, weight decay, warmup schedule, dropout, batch size, and first/second order moment coefficients as the heavily-tuned default first order baseline optimizers. Additional gains could be achieved if one wanted to optimize extensively these hyperparameters.

II. PROBLEM FORMULATION AND RELATED WORK

We focus on supervised learning tasks where the goal is to solve a non-convex stochastic optimization problem of the form:

$$\min_{\theta} \mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N l_i(x_i, y_i; \theta), \quad (1)$$

where $\theta \in \mathbb{R}^d$ denotes the model parameters, $l_i(x_i, y_i; \theta)$ is the loss function, (x_i, y_i) is the paired input data and its corresponding ground truth label, and N is the total

number of data points in the training dataset. Furthermore, we denote the gradient of the loss w.r.t. model parameters as $\mathbf{g} = \frac{1}{N_B} \sum_{i=1}^{N_B} \frac{\partial l_i}{\partial \theta}$, and the corresponding second derivative (i.e., Hessian) as $\mathbf{H} = \frac{1}{N_B} \sum_{i=1}^{N_B} \frac{\partial^2 l_i}{\partial \theta^2}$, where N_B is the size of one mini-batch.

Solving Eq. 1 for a real learning problem (and not a simple model) is a very challenging task. Despite years of research, we have not yet been able to resolve several seemingly ad-hoc tricks that are required to converge (hopefully) to a *good* solution. Next, we briefly discuss the different popular optimization methods proposed in recent years to address the challenges associated with solving Eq. 1. This is by no means a comprehensive review, and we refer the interested reader to [8] for a thorough review.

A. Adaptive First Order Methods

Due to their simplicity and effectiveness, first order optimization methods [20, 29, 34, 41, 49, 71] have become the de-facto algorithms used in deep learning. There are multiple variations, but these methods can be represented using the following general update formula:

$$\theta_{t+1} = \theta_t - \eta_t m_t / v_t, \quad (2)$$

where η_t is the learning rate, and m_t , and v_t denote the so called first and second moment terms, respectively. A simple and popular update method is SGD, originally proposed in 1951 as a root-solving algorithm [49]:

$$m_t = \beta m_{t-1} + (1 - \beta) \mathbf{g}_t \quad \text{and} \quad v_t \equiv 1. \quad (3)$$

Here, \mathbf{g}_t is the gradient of a mini-batch at t -th iteration and β is the momentum hyperparameter.

Using SGD to solve Eq. 1 is often very challenging, as the convergence of the iterative formulae in Eq. 2 is very sensitive to the right choice of the learning rate, its decay schedule, and the momentum parameter. To address this, several methods have been proposed to take into account the knowledge of the geometry of the data by scaling gradient coordinates, using the past gradient information. This can be viewed in one of two equivalent ways: either as automatically adjusting the learning rate in Eq. 2; or as an adaptive *preconditioning* of the gradient. One notable method is Adagrad [20, 37], which accumulates all the gradients from the first iteration and applies the square

root of the result to precondition the current gradient. The update formulae in this case become¹:

$$m_t = \mathbf{g}_t \quad \text{and} \quad v_t = \sqrt{\sum_{i=1}^t \mathbf{g}_i \mathbf{g}_i}. \quad (4)$$

While Adagrad works well for sparse settings, its performance significantly degrades for dense settings, which is the case for many machine learning tasks. In particular, this stems from the accumulation of all previous gradients for the preconditioner Eq. 4. This results in a monotonic increase in the magnitude of the second moment, v_t , which effectively translates into a rapid decay of the learning rate. To address this, several methods have been proposed where the intuition is to limit the accumulation to a small window of past iterations, and in particular exponentially reduce the weight of earlier iterations. Notable works incorporating this method are RMSProp, ADADelta, and Adam [29, 56, 71]. In particular, for Adam [29], the two moments for the update rule are the following:

$$m_t = \frac{(1 - \beta_1) \sum_{i=1}^t \beta_1^{t-i} \mathbf{g}_i}{1 - \beta_1^t}, \quad (5)$$

$$v_t = \sqrt{\frac{(1 - \beta_2) \sum_{i=1}^t \beta_2^{t-i} \mathbf{g}_i \mathbf{g}_i}{1 - \beta_2^t}},$$

where $0 < \beta_1, \beta_2 < 1$ are two hyperparameters sometimes referred to as first and second moment coefficients. In particular, note that the sum over past gradients is scaled by β_2 which exponentially reduces the contribution of early gradients. A summary of the different m_t and v_t used by common first-order optimizers is given in Table 1. A notable variant here is AdamW [34], which shows that decoupling weight decay from the update equation of Adam can lead to a noticeable performance improvement. Recently, AdamW has become the preferred optimizer for NLP tasks, and in particular for training transformers [57]. There are also many other variants of adaptive first order methods [12, 33, 54, 55, 73].

Despite all these attempts, it is still not clear which optimizer should work for a *new* learning task/model. This is in fact one of the main baffling practical issues in machine learning, and one for which theory has little to say. For example, SGD is currently the best performing optimizer for some CV tasks. That is, using other variants such as AdamW leads to significantly worse

Table 1: Summary of the first and second moments used in different optimization algorithms for updating model parameters ($\theta_{t+1} = \theta_t - \eta m_t / v_t$). Here β_1 and β_2 are first and second moment hyperparameters.

Optimizer	m_t	v_t
SGD [49]	$\beta_1 m_{t-1} + (1 - \beta_1) \mathbf{g}_t$	1
Adagrad [20]	\mathbf{g}_t	$\sqrt{\sum_{i=1}^t \mathbf{g}_i \mathbf{g}_i}$
Adam [29]	$\frac{(1 - \beta_1) \sum_{i=1}^t \beta_1^{t-i} \mathbf{g}_i}{1 - \beta_1^t}$	$\sqrt{\frac{(1 - \beta_2) \sum_{i=1}^t \beta_2^{t-i} \mathbf{g}_i \mathbf{g}_i}{1 - \beta_2^t}}$
RMSProp [56]	\mathbf{g}_t	$\sqrt{\beta_2 v_{t-1}^2 + (1 - \beta_2) \mathbf{g}_t \mathbf{g}_t}$
ADAHESIAN	$\frac{(1 - \beta_1) \sum_{i=1}^t \beta_1^{t-i} \mathbf{g}_i}{1 - \beta_1^t}$	$\sqrt{\frac{(1 - \beta_2) \sum_{i=1}^t \beta_2^{t-i} \mathbf{D}_i^{(s)} \mathbf{D}_i^{(s)}}{1 - \beta_2^t}}$

generalization performance. However, for NLP tasks, AdamW has the best performance by a large margin as compared to SGD. The point here is that even the choice of the optimizer has effectively become a hyperparameter.

B. Second Order Methods

Second order methods are among the most powerful optimization methods that have been designed, and there have been several attempts to use their many advantages for training NNs. Second order methods are designed to address *ill-conditioned* optimization problems by automatically rotating and rescaling the gradient. This allows one to choose a better descent direction, and to adjust automatically the learning rate for each parameter. There have also been multiple theoretical studies showing better convergence rate of second order based methods [2, 3, 6, 11, 13, 14, 50, 61–63, 65, 69]. In particular, second order methods can guarantee convergence to second order critical points, while the vast majority of first order methods lack such guarantees. For example, theoretically it has been shown that some first order methods can only converge to an approximate second order critical point [21, 28, 32, 48].

Newton’s method is a classical second order method where one solves a linear system, essentially meaning that the inverse of the local Hessian is used at every iteration to precondition the gradient vector.² One major challenge with this approach is that it can be expensive to solve the linear system, naïvely requiring cubic computational complexity, not including the cost of forming the Hessian

¹Throughout the paper, without further notification, for two vectors, e.g., a and b , we use both ab and $a \odot b$ to denote the element-wise product, and $\langle a, b \rangle$ denotes the inner product.

²To be clear, when we refer to computing an inverse, we mean that we use a numerical method that performs a linear equation solve that effectively amounts to working with the inverse implicitly. Of course, one would never actually compute the Hessian inverse explicitly.

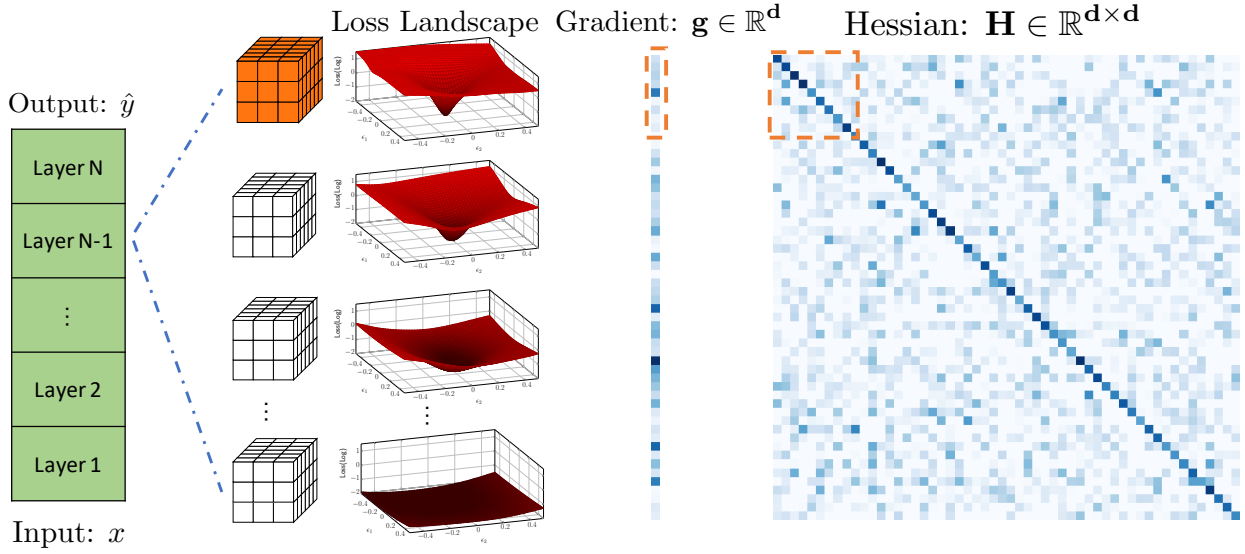


Fig. 2: A simple model with N layers (first column); with the convolutional blocks of the $N-1$ layer shown (second column); and the loss landscape of each block (third column), which can be calculated by perturbing the convolutions's parameters in two different eigendirections. (See [66] for details of how to construct loss landscape.) Note the different loss landscape topologies. First order methods do not explicitly capture this difference. The entries (3D tensors) colored in orange show the components used for calculating the spatial average of Hessian. The part of the gradient (fourth panel) highlighted in the orange box is the corresponding gradient of the orange convolution kernel; and the part of the Hessian diagonal (fifth panel) highlighted in the orange box is used to compute the spatial average.

itself and the corresponding quadratic memory complexity. However, the overhead of such a naïve implementation can be improved by using so-called matrix free methods, where the Hessian matrix is never explicitly formed (addressing quadratic memory cost), and its inverse is approximately and only implicitly applied (addressing the cubic computational complexity).

One seminal work here is the limited memory BFGS (LBFGS) [10] method which has a desirable linear computational and memory complexity. This approach approximates the Hessian as a series sum of first order information from prior iterations. As such, these approaches that do not directly use the Hessian operator are referred to as *Quasi-Newton* methods. While this approach works well for many optimization problems, it does not work well for many machine learning problems. One reason for this is that LBFGS method requires full batch gradients, as stochastic gradients can lead to drastic errors in the approximation [7]. This is one of the main challenges with Quasi-Newton methods applied to machine learning problems [4]. Other approaches such as approximating the Hessian as the Kronecker product of vectors have also been explored [36].

There has also been work on enhancing first order methods by incorporating the Fisher information matrix [22]. The main idea is to use the Fisher information instead of the squared norm of the gradient. A naïve use of Fisher

information has computational and memory overhead, but it is possible to also approximate the Fisher information matrix using low rank decomposition [22].

Another line of work has been to incorporate automatically the Hessian operator itself, instead of approximating it using first order information. A work here is [53] which uses Gauss-Newton Hessian diagonal to adjust adaptively the learning rate. The work of [62] also directly incorporates the Hessian using a trust region method.

While the above approaches are very interesting and result in a good performance for simple models, they do not achieve comparable results for more complex NN architectures. One of the reasons that second order methods have not been successful yet for machine learning, as opposed to other domains such as scientific computing, is due to the stochastic nature of the problem. Such stochastic noise leads to an erroneous approximation of the Hessian, leading to suboptimal descent directions. SGD is more robust to such noise since we can efficiently incorporate moving averages and momentum. Ideally, if there was a way to apply the same moving average method to the Hessian, then that would help smooth out local curvature noise to get a better approximation to the non-noisy curvature of the loss landscape. However, such an approximation is challenging since the Hessian is a matrix that cannot be explicitly formed to be averaged, whereas it is easy to form the gradient vector.

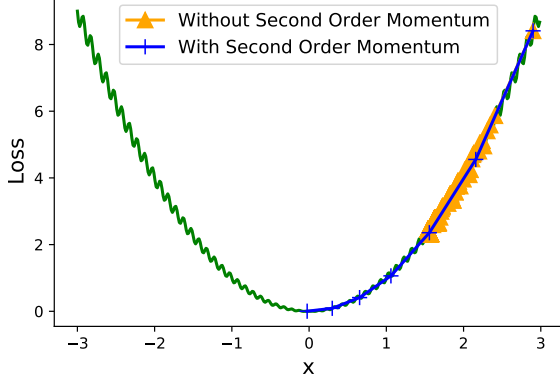


Fig. 3: Local versus global curvature. Illustration of the local curvature which can be noisy, and the global curvature with a simple 1D problem $f(x) = x^2 + 0.1x \sin(20\pi x)$. Using the exponential moving average of Eq. 12 is key to avoid the misleading local curvature information. To demonstrate this we test ADAHESSIAN without moving average (orange trajectory) which does not converge even after 1000 iterations. On the other hand, ADAHESSIAN converges in 7 iterations with the moving average enabled.

As we show below, ADAHESSIAN addresses this problem by incorporating the Hutchinson’s method along with spatial averaging to reduce the impact of the stochastic noise. The result exceeds the performance of all the above methods for machine learning tasks. Next, we formally introduce the ADAHESSIAN algorithm.

III. METHODOLOGICAL APPROACH

Here, we first provide the formulation for the full Newton method in Section III-A. Then, we describe the three components of ADAHESSIAN, namely Hessian diagonal approximation (Section III-B), spatial averaging (Section III-C), and Hessian momentum (Section III-D). Finally, we discuss the overall formulation of ADAHESSIAN in Section III-E.

A. A General Hessian Based Descent Direction

For the loss function $f(w) : \mathbb{R}^d \rightarrow \mathbb{R}$, let us denote the corresponding gradient and Hessian of $f(w_t)$ at iteration t as \mathbf{g}_t , and \mathbf{H}_t , respectively.³ A general descent direction can then be written as follows for a positive-definite Hessian:

$$\Delta w_t = \mathbf{H}_t^{-k} \mathbf{g}_t, \quad \text{where } \mathbf{H}_t^{-k} = U_t^T \Lambda_t^{-k} U_t. \quad (6)$$

Here, we refer to $0 \leq k \leq 1$ as *Hessian power*, and $U_t^T \Lambda_t U_t$ is the eigendecomposition of \mathbf{H}_t . Note that

³Without confusion, we use the same gradient and Hessian notations for $f(w)$ and $\mathcal{L}(\theta)$. Furthermore, when there is no confusion we will drop subscript t .

for $k = 0$, we recover the gradient descent method; and for $k = 1$, we recover the Newton method. In our empirical tests we consider non-convex machine learning problems, but we provide a standard convergence behaviour of Eq. 6 in Appendix A for a simple strongly convex and strictly smooth function $f(w)$. (We emphasize that the proof is very standard and we are only including it for completeness.)

The basic idea of Hessian based methods is to *pre-condition* the gradient with the \mathbf{H}^{-k} and use $\mathbf{H}^{-k} \mathbf{g}$ for the update direction, instead of using the *bare* gradient \mathbf{g} vector. The preconditioner automatically rotates and rescales the gradient vector. This is important since the loss landscape curvature is generally different across different directions/layers and since these directions need not correspond to the canonical axes. This is illustrated in Figure 2, where we show a 2D schematic plot of the loss landscape for different convolution channels [66]. Each channel can have a different loss landscape topology. For example, the last channel has a much flatter loss landscape, as compared to other layers. As a result, it is preferable to take a larger step size for the last channel than for the first channel, which has a very “sharp” loss landscape. Problems that exhibit this behaviour are *ill-conditioned*. The role of the Hessian is to automatically normalize this ill-conditionedness by stretching and contracting different directions to accommodate for the curvature differences (full Newton method also rotates the gradient vector along with adjusting the step size).

However, there are two major problems with this approach. The first problem is that a naïve use of the Hessian preconditioner comes at the prohibitively high cost of applying Hessian inverse to the gradient vector at every iteration ($\mathbf{H}^{-k} \mathbf{g}$ term). The second and more challenging problem is that local Hessian (curvature) information can be very misleading for a noisy loss landscape. A simple example is illustrated in Figure 3, where we plot a simple parabola with a small sinusoidal noise as the loss landscape (shown in green). As one can see, the local Hessian (curvature) information is completely misleading, as it computes the curvature of the sinusoidal noise instead of global Hessian information for the parabola. Applying such misleading information as the preconditioner would actually result in very small steps to converge to one of the many local minima created by the sinusoidal noise. The same problem exists for the gradient as well, but that can be alleviated by using gradient momentum instead of local gradient information. However, as mentioned before it is computationally infeasible to compute (naïvely) a Hessian momentum. The reason is that we cannot form

the Hessian matrix and average it throughout different iterations, as such an approach has quadratic memory complexity in the number of parameters along with a prohibitive computational cost. However, one could use Randomized Numerical Linear Algebra to get a sketch of the Hessian matrix [23, 66, 67]. In particular, we show how this can be done to approximate the Hessian diagonal. However, as we discuss next, both problems can be resolved by using Hessian diagonal instead of the full Hessian.

B. Hessian Diagonal Approximation

To address the issue that applying the inverse Hessian to the gradient vector at every iteration is computationally infeasible, one could use an inexact Newton method, where an approximate Hessian operator is used instead of the full Hessian [6, 16, 63, 64, 69]. The most simple and computationally efficient approach is to approximate the Hessian as a diagonal operator in Eq. 6:

$$\Delta w = \text{diag}(\mathbf{H})^{-k} \mathbf{g}, \quad (7)$$

where $\text{diag}(\mathbf{H})$ is the Hessian diagonal, which we denote as \mathbf{D} .⁴ We show that using Eq. 7 has the same convergence rate as using Eq. 6 for simple strongly convex and strictly smooth function $f(w)$ (see Appendix B). Note that we only include the proof for completeness, and our algorithm ADAHESSIAN can be applied for general machine learning problems.

The Hessian diagonal \mathbf{D} can be efficiently computed using the Hutchinson’s method. The two techniques we use for this approximation are: (i) a Hessian-free method [67]; and (ii) a randomized numerical linear algebra (RandNLA) method [5, Figure 1]. In particular, the Hessian-free method is an oracle to compute the multiplication between the Hessian matrix \mathbf{H} with a random vector z , i.e.,

$$\frac{\partial \mathbf{g}^T z}{\partial \theta} = \frac{\partial \mathbf{g}^T}{\partial \theta} z + \mathbf{g}^T \frac{\partial z}{\partial \theta} = \frac{\partial \mathbf{g}^T}{\partial \theta} z = \mathbf{H}z. \quad (8)$$

Here, the first equality is the chain rule, and the second equality is since z is independent of θ . Eq. 8 effectively allows us to compute the Hessian times a vector z , without having to form explicitly the Hessian, by backpropotating the $\mathbf{g}^T z$ term. This has the same cost as ordinary gradient backpropogation [67]. Then, with the Hessian matvec oracle, one can compute the Hessian diagonal using Hutchinson’s method:

$$\mathbf{D} = \text{diag}(\mathbf{H}) = \mathbb{E}[z \odot (\mathbf{H}z)], \quad (9)$$

⁴Note that \mathbf{D} can be viewed as a vector, in which case $\mathbf{D}^{-k} \mathbf{g}$ is an element-wise product of vectors. Without clarification, \mathbf{D} is treated as a vector for the rest of the paper.

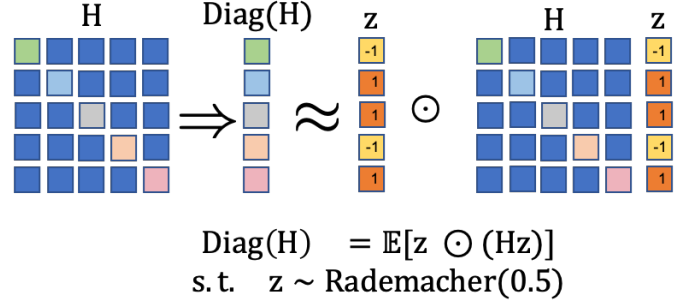


Fig. 4: Illustration of the diagonal Hessian estimation with Hutchinson’s method.

where z is a random vector with Rademacher distribution, and $\mathbf{H}z$ is computed by the Hessian matvec oracle given in Eq. 8. This process is illustrated in Figure 4. It can be proved that the expectation of $z \odot (\mathbf{H}z)$ is the Hessian diagonal [5].

Another important advantage, besides computational efficiency, of using the Hessian diagonal is that we can compute its moving average to resolve the local noisy Hessian as mentioned at the end of Section III-A. This allows us to smooth out noisy local curvature information, and to obtain estimates that use global Hessian information instead. We incorporate both spatial averaging and momentum (temporal averaging) to smooth out this noisy Hessian estimate as described next.

C. Spatial Averaging

The Hessian diagonal can vary significantly for each single parameter dimension of the problem. We found it helpful to perform spatial averaging of Hessian diagonal and use the average to smooth out spatial variations. For example, for a convolutional layer, each convolution parameter can have a very different Hessian diagonal. In ADAHESSIAN we compute the average of the Hessian diagonal for each convolution kernel (3×3) as illustrated in Figure 5. Mathematically, we perform a simple spatial averaging on the Hessian diagonal as follows:

$$\mathbf{D}^{(s)}[ib+j] = \frac{\sum_{k=1}^b \mathbf{D}[ib+k]}{b}, \text{ for } 1 \leq j \leq b, 0 \leq i \leq \frac{d}{b}-1, \quad (10)$$

where $\mathbf{D} \in \mathbb{R}^d$ is the Hessian diagonal, $\mathbf{D}^{(s)} \in \mathbb{R}^d$ is the spatially averaged Hessian diagonal, $\mathbf{D}[i]$ ($\mathbf{D}^{(s)}[i]$) refers to the i -th element of \mathbf{D} ($\mathbf{D}^{(s)}$), b is the spatial average block size, and d is the number of model parameters divisible by b . We show that replacing \mathbf{D} in Eq. 7 by $\mathbf{D}^{(s)}$ in Eq. 10, the update direction has the same convergence rate as using Eq. 6 for simple strongly convex and strictly smooth function $f(w)$ (see Appendix C). Note

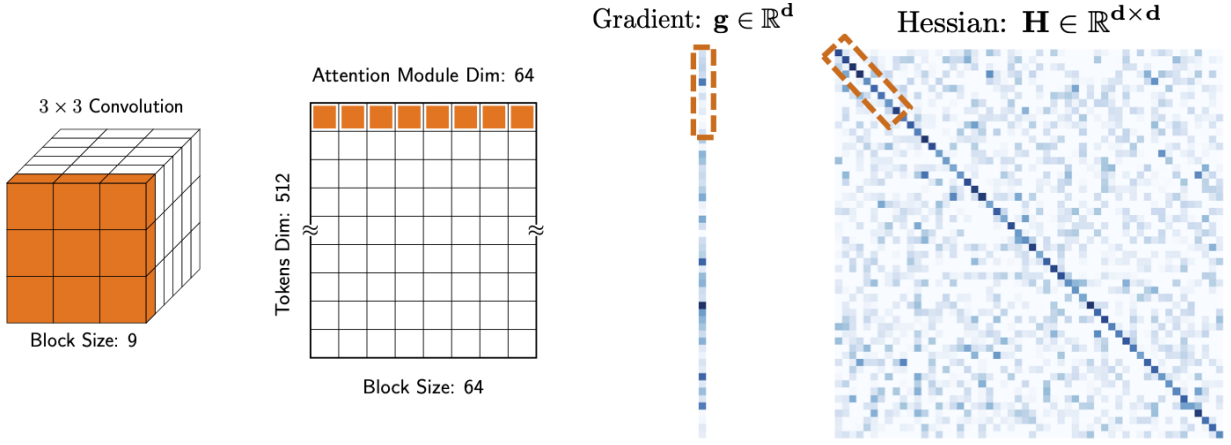


Fig. 5: Illustration of the block size used to average the Hessian diagonal to smooth spatial variations. For a convolution layer, we average each channel (groups of 9 parameters); and for multi-head attention, we average consecutive elements along the rows (attention dimension). We found that using block averaging helps, although ADAHESSIAN is not very sensitive to this hyperparameter as illustrated in Table 7.

that we only include the proof for completeness, and our algorithm ADAHESSIAN can be applied for general machine learning problems.

Figure 5 provides illustration of spatial averaging for both convolutional and matrix kernels. In general, the block size b is a hyperparameter that can be tuned for different tasks. While this is a new hyperparameter that can help the performance, the performance of ADAHESSIAN is not very sensitive to it (we provide sensitivity results in Section V-A).

Next we describe momentum which is another useful method to smooth out Hessian noise over different iterations.

Algorithm 1: ADAHESSIAN

Require: Initial Parameter: θ_0
Require: Learning rate: η
Require: Exponential decay rates: β_1, β_2
Require: Block size: b
Require: Hessian Power: k
Set: $m_0 = 0, v_0 = 0$
for $t = 1, 2, \dots$ **do** // Training Iterations
 $\mathbf{g}_t \leftarrow$ current step gradient
 $\mathbf{D}_t \leftarrow$ current step estimated diagonal Hessian
 Compute $\mathbf{D}_t^{(s)}$ based on Eq. 10
 Update $\bar{\mathbf{D}}_t$ based on Eq. 11
 Update m_t, v_t based on Eq. 12
 $\theta_t = \theta_{t-1} - \eta m_t / v_t$

D. Hessian Momentum

We can easily apply momentum to Hessian diagonal since it is a vector instead of a quadratically large matrix.

This enables us to adopt momentum for Hessian diagonal in ADAHESSIAN. More specifically, let $\bar{\mathbf{D}}_t$ denote the Hessian diagonal with momentum that is calculated as:

$$\bar{\mathbf{D}}_t = \sqrt{\frac{(1 - \beta_2) \sum_{i=1}^t \beta_2^{t-i} \mathbf{D}_i^{(s)} \mathbf{D}_i^{(s)}}{1 - \beta_2^t}}, \quad (11)$$

where $\mathbf{D}^{(s)}$ is the spatially averaged Hessian diagonal (defined in Eq. 10), and $0 < \beta_2 < 1$ is the second moment hyperparameter. Note that this is exactly the same as the momentum term in Adam [29] or RMSProp [56] except that we are using the spatial averaging Hessian diagonal instead of the gradient.

To illustrate the importance of Hessian momentum, we provide a simple example in 1D by considering $f(x) = x^2 + 0.1x \sin(20\pi x)$, as shown in Figure 3. It can be clearly seen that the method without the second order momentum gets trapped at a local minima even with more than 1000 iterations (orange trajectory). On the contrary, the optimization converges within 7 iterations with Hessian momentum (blue trajectory). (While this example is over-simplified in certain ways, we are using it here only to convey the importance of momentum.)

E. AdaHessian

To summarize, instead of only applying momentum for gradient, ADAHESSIAN uses *spatial averaging* and *Hessian momentum* to smooth out local variations in Hessian diagonal. More specifically, the first and second

order moments (m_t and v_t) for ADAHESSIAN are computed as follows:

$$m_t = \frac{(1 - \beta_1) \sum_{i=1}^t \beta_1^{t-i} \mathbf{g}_i}{1 - \beta_1},$$

$$v_t = (\bar{D}_t)^k = \left(\sqrt{\frac{(1 - \beta_2) \sum_{i=1}^t \beta_2^{t-i} \mathbf{D}_i^{(s)} \mathbf{D}_i^{(s)}}{1 - \beta_2}} \right)^k, \quad (12)$$

where $0 < \beta_1, \beta_2 < 1$ are the first and second moment hyperparameters that are also used in Adam. Note that Adam uses the same formulation except that the spatial averaging Hessian diagonal $\mathbf{D}_i^{(s)}$ is replaced with gradient.

The main overhead of ADAHESSIAN is the Hutchinson’s method to approximate Hessian diagonal, \mathbf{D} . We use one Hutchinson step per iteration to approximate the Hessian diagonal (i.e., one random Rademacher vector z in Eq. 9). The cost of this estimation is one Hessian matvec (to compute $H z$), which is equivalent to one gradient backpropagation [66, 67].

Also note that it is possible to get a more accurate approximation to Hessian diagonal by using more Hutchinson steps per iteration. However, we found that one step per iteration performs well in practice since the multiple calculations could be performed as Hessian momentum (Section III-D). In fact, as we discuss in Section V-B, it is possible to skip the Hutchinson calculation for few iterations to reduce further its computational overhead, without significant impact on final accuracy.

IV. RESULTS

A. Experiment Setup

One of the problems with several formerly proposed optimization methods is that the methods were originally tested with very simple models on very few tasks. When those methods were later tested by the community on more complex models the results were often worse than popular optimization methods. To avoid such a scenario, we extensively test ADAHESSIAN on a wide range of learning tasks, including image classification, neural machine translation (NMT), language modeling (LM), and recommendation system (RecSys). We compare the ADAHESSIAN performance with SGD, Adam, AdamW [34], and Adagrad. Moreover, to enable a fair comparison we will use the same β_1 and β_2 parameters in ADAHESSIAN as in Adam/AdamW for each task, even though those default values may favor Adam (or AdamW) and disfavor ADAHESSIAN. Furthermore, we will use the exact same weight decay and learning rate schedule

Table 2: Results of ResNet20/32 on Cifar10 (left two columns) and ResNet18 on ImageNet (last column). On Cifar10: Adam performs consistently worse than SGD; AdamW has slightly worse performance than SGD; and ADAHESSIAN outperforms AdamW and even gets accuracy comparable to SGD. On ImageNet: ADAHESSIAN has significantly better accuracy than Adam (5.53%), AdamW (2.67%), and has similar performance to SGD.

Dataset	Cifar10		ImageNet
	ResNet20	ResNet32	ResNet18
SGD [51]	92.08 \pm 0.08	93.14 \pm 0.10	70.03
Adam [29]	90.33 \pm 0.13	91.63 \pm 0.10	64.53
AdamW [34]	91.97 \pm 0.15	92.72 \pm 0.20	67.41
ADAHESSIAN	92.13 \pm 0.18	93.08 \pm 0.10	70.08

in ADAHESSIAN as that used by other optimizers. Below we briefly explain each of the learning tasks tested.

a) Image Classification: We experiment on both Cifar10 (using ResNet20/32) and ImageNet (using ResNet18) datasets. Cifar10 consists of 50k training images and 10k testing images. ImageNet has 1.2M training images and 50k validation images. We follow the settings described in [25] for training. We run each experiment 5 times on Cifar10 and report the mean and standard deviation of the results.

b) Neural Machine Translation (NMT): We use IWSLT14 German-to-English (De-En) and WMT14 English-to-German (En-De) datasets. Transformer base architecture is used for WMT14 (4.5M sentence pairs), and small architecture is used for IWSLT14 (0.16M sentence pairs). We follow the settings reported in [43] and use pre-normalization described in [59]. The length penalty is set to 0.6/1.0 and the beam size is set to 4/5 for WMT/IWSLT [42]. We report the average results of the last 10/5 checkpoints respectively. For NMT, BLEU score is used [44]. In particular, we report tokenized case-sensitive BLEU on WMT14 En-De and case-insensitive BLEU IWSLT14 De-En. Furthermore, we use AdamW for this task instead of Adam since the former is the standard optimizer (Adam consistently scores lower).

c) Language Modeling: We use PTB [39] and Wikitext-103 [38] datasets, which contain 0.93M and 100M tokens, respectively. Following [35], a three-layer tensorized transformer core-1 for PTB and a six-layer tensorized transformer core-1 for Wikitext-103 are used in the experiments. We apply the multi-linear attention mechanism with masking and report the perplexity (PPL) on the test set with the best validation model.

d) Natural Language Understanding: We use the GLUE task [58] to evaluate the fine-tuning performance of SqueezeBERT [27]. More specifically, we use 8

Table 3: NMT performance (BLEU) on IWSLT14 De-En and WMT14 En-De testsets (higher is better). Unlike in Table 2, SGD has significantly worse results than AdamW. Note that ADAHESSIAN outperforms the default and heavily tuned optimizer AdamW by 0.13 and 0.33 on IWSLT14 and WMT14, which is significant for this task.

Model	IWSLT14	WMT14
	small	base
SGD	28.57 \pm .15	26.04
AdamW [34]	35.66 \pm .11	28.19
ADAHESSIAN	35.79 \pm .06	28.52

different tasks in GLUE and report and final average performance on the validation dataset.

e) Recommendation System: The Criteo Ad Kaggle dataset contains approximately 45 million samples over 7 days. We follow the standard setting and use the first 6 days as the training set and the last day as the test set. Furthermore, we use DLRM, a novel recommendation model that has been recently released by Facebook [40]. The testing metric for Recommendation Systems is Click Through Rate (CTR), measured on training and test sets.

We refer the interested reader to Appendix D for more detailed experimental settings. Next we report the experimental results on each of these tasks.

B. Image Classification

The results on Cifar10 are shown in Table 2. First, note the significantly worse performance of Adam, as compared to SGD even on this simple image classification dataset. Particularly, Adam has 1.75%/1.51% lower accuracy for ResNet20/32 than SGD. AdamW achieves better results than Adam, but its performance is still slightly worse than SGD. However, ADAHESSIAN achieves significantly better results as compared to Adam (1.80%/1.45% for ResNet20/32), even though we use the same β_1 and β_2 parameters in ADAHESSIAN as in Adam. That is, we did not tune these two hyperparameters, even though tuning them could potentially lead to even better performance.⁵ Compared with SGD, ADAHESSIAN achieves comparable accuracy for both ResNet20 (0.05% higher) and ResNet32 (0.06% lower). The training and testing curves of different optimizers for ResNet20/32 on Cifar10 are shown in Figure A.7.

Next, we use the best learning rate obtained by training ResNet20/32 on Cifar10 to optimize ResNet18

⁵In fact, in Table 8 we achieve 92.40 for ResNet20 which is higher than what we report in Table 2. This is to emphasize that we only tuned learning rate in Table 2. Still ADAHESSIAN achieves significantly better results than Adam.

Table 4: LM performance (PPL) on PTB and Wikitext-103 test datasets (lower is better). The PPL of ADAHESSIAN is 2.7 and 1.0 lower than that of AdamW.

Model	PTB	Wikitext-103
	Three-Layer	Six-Layer
SGD	59.9 \pm 3.0	78.5
AdamW [34]	54.2 \pm 1.6	20.9
ADAHESSIAN	51.5 \pm 1.2	19.9

on ImageNet for all four optimizers. We try two different learning rate schedules for all four optimizers, and we use the one with the better result. The two learning rate schedules are quite standard, i.e., the step decay schedule and the plateau based schedule [46]. The final result is reported in Table 2. Again note that the final performances of Adam and AdamW are much worse than that of SGD and ADAHESSIAN. We plot the training and testing curve in Figure A.8.

It is worthwhile to note that our learning rate tuning is performed at an academic scale, but ADAHESSIAN still significantly exceeds other adaptive methods and reaches the same performance level as SGD which has been tuned at the industrial scale.

C. Neural Machine Translation

We use BLEU [44] as the evaluation metric for NMT. Following standard practice, we measure tokenized case-sensitive BLEU and case-insensitive BLEU for WMT14 En-De and IWSLT14 De-En, respectively. For a fair comparison, we do not include other external datasets.

The NMT results are shown in Table 3. The first interesting observation is that here SGD performs much worse than AdamW (which is opposite to its behaviour for image classification problems where SGD has superior performance; see Appendix IV-B). As pointed out in the introduction, even the choice of the optimizer has become another hyperparameter. In particular, note that the BLEU scores of SGD are 7.09 and 2.15 lower than AdamW on IWSLT14 and WMT14, which is quite significant. Similar observations about SGD were also reported in [72].

Despite this, ADAHESSIAN achieves state-of-the-art performance for NMT with transformers. In particular, ADAHESSIAN outperforms AdamW by 0.13 BLEU score on IWSLT14. Furthermore, the accuracy of ADAHESSIAN on WMT14 is 28.52, which is 0.33 higher than that of AdamW. We also plot the training losses of AdamW and ADAHESSIAN on IWSLT14/WMT14 in Figure A.9. As one can see, ADAHESSIAN consistently achieves lower

Table 5: Comparison of AdamW and ADAHESSIAN for SqueezeBERT on the development set of the GLUE benchmark. As can be seen, the average performance of ADAHESSIAN is 0.41 higher as compared to AdamW. The result of AdamW⁺ is directly from [27] and the result of AdamW* is reproduced by us.

	RTE	MPRC	STS-B	SST-2	QNLI	QQP	MNLI-m	MNLI-mm	Avg.
AdamW ⁺ [27]	71.8	89.8	89.4	92.0	90.5	89.4	82.9	82.3	86.01
AdamW*	79.06	90.69	90.00	91.28	90.30	89.49	82.61	81.84	86.91
ADAHESSIAN	80.14	91.94	90.59	91.17	89.97	89.33	82.78	82.62	87.32

training loss. These improvements are quite significant for NMT, and importantly these are achieved even though ADAHESSIAN directly uses the same β_1 and β_2 , as well as the same number of warmup iterations as in AdamW.

D. Language Modeling

We report the language modeling results in Table 4, using the tensorized transformer proposed in [35]. Similar to NMT, note that the perplexity (PPL) of SGD is more than 57 points worse than AdamW on Wikitext-103. That is, similar to the NMT task, SGD performs worse than AdamW. However, ADAHESSIAN achieves more than 1.8/1.0 better PPL than that of AdamW on PTB/Wikitext-103, respectively.

We also show the detailed training loss curves in Figure A.10. ADAHESSIAN achieves consistently lower loss values than AdamW throughout the training process on both PTB and Wikitext-103. Similar to NMT, the β_1/β_2 as well as the warmup phase of ADAHESSIAN are kept the same as AdamW.

E. Natural Language Understanding

We report the NLU results in Table 5, using the SqueezeBERT model [26] tested on GLUE datasets [58]. As can be seen, ADAHESSIAN has better performance than AdamW on 5 out of 8 tasks. Particularly, on RTE and MPRC, ADAHESSIAN achieves more than 1 point as compared to AdamW. On average, ADAHESSIAN outperforms AdamW by 0.41 points. Note that similar to NMT and LM, except learning rate and block size, ADAHESSIAN directly uses the same hyperparameters as AdamW. Interestingly, note that these results are better than those reported in SqueezeBERT [27], even though we only change the optimizer to ADAHESSIAN instead of AdamW.

F. Recommendation System

We solely focus on modern recommendation systems, and in particular on the DLRM model widely adopted in industry [40]. These systems include a large embedding layer followed by a series of dense FC layers. In training,

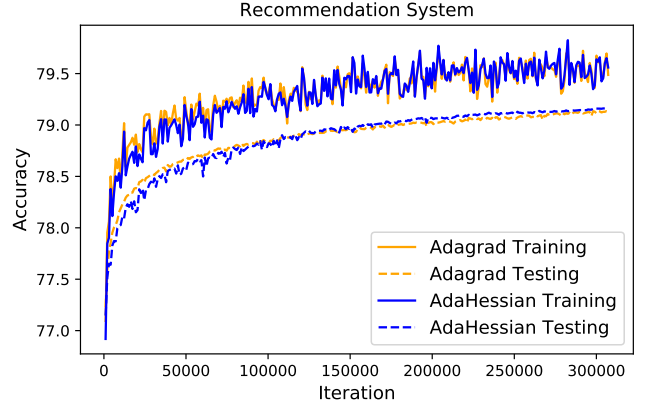


Fig. 6: Training and Testing Accuracy curves of Adagrad and ADAHESSIAN on Criteo Ad Kaggle dataset. As can be seen, the test accuracy of ADAHESSIAN is better (0.032%) than that of Adagrad. This is quite significant for this task.

a sparse set of rows of the embedding layer is used and only those rows are updated. These rows do change from one iteration to the next. For such a sparse setting, we use Adagrad to update the embedding table, and we use ADAHESSIAN to update the rest of the FC network in the experiments. (Pytorch currently does not support second order backpropagation for the sparse gradient to the embedding.) ADAHESSIAN uses the same hyperparameters for updating the embedding table as in the Adagrad experiment without tuning. The training and testing accuracy curves are reported in Figure 6. The testing accuracy of ADAHESSIAN is 79.167%, which is 0.032% higher than Adagrad. It should be noted that this is a quite significant accuracy increase for Recommendation Systems [60].

V. DISCUSSION

As reported in the previous section, ADAHESSIAN achieves state-of-the-art performance on a wide range of tasks. Two important issues are the sensitivity of ADAHESSIAN to the hyperparameters of learning rate and block size. This is discussed next.

Table 6: Robustness of AdamW and ADAHESSIAN to the learning rate on IWSLT14. We scale the base learning rate used in Section IV-C. As can be seen, ADAHESSIAN is much more robust to large learning rate variability as compared to AdamW.

LR Scaling	0.5	1	2	3	4	5	6	10
AdamW	35.42 \pm .09	35.66 \pm .11	35.37 \pm .07	35.18 \pm .07	34.79 \pm .15	14.41 \pm 13.25	0.41 \pm .32	Diverge
ADAHESSIAN	35.33 \pm .10	35.79 \pm .06	35.21 \pm .14	34.74 \pm .10	34.19 \pm .06	33.78 \pm .14	32.70 \pm .10	32.48 \pm .83

Table 7: Block Size effect of ADAHESSIAN on IWSLT14. With various block sizes, the performance of ADAHESSIAN is very stable and no worse than that of AdamW (35.66 \pm .11).

Block Size	1	2	4	8	16	32	64	128
ADAHESSIAN	35.67 \pm .10	35.66 \pm .07	35.78 \pm .07	35.77 \pm .08	35.67 \pm .08	35.79 \pm .06	35.72 \pm .06	35.67 \pm .11

A. Learning Rate and Block Size Effects

Here, we explore the effects of the learning rate and block size b on ADAHESSIAN. We first start with the effect of learning rate, and test the performance of ADAHESSIAN and AdamW with different learning rates. The results are reported in Table 6 for IWSLT14 dataset, where we scale the original learning rate with a constant factor, ranging from 0.5 to 20 (the original learning rate is the same as in Section IV-C). It can be seen that ADAHESSIAN is more robust to the large learning rates. Even with 10 \times learning rate scaling, ADAHESSIAN still achieves 32.48 BLEU score, while AdamW diverges even with 6 \times learning rate scaling. This is a very desirable property of ADAHESSIAN, as it results in reasonable performance for such a wide range of learning rates.

We also test the effect of the spatial averaging block size (parameter b in Eq. 10). As a reminder, this parameter is used for spatially averaging the Hessian diagonal as illustrated in Figure 5. The sensitivity results are shown in Table 7 where we vary the block size from 1 to 128. While the best performance is achieved for the block size of 32, the performance variation for other block sizes is rather small. Moreover, all the results are still no worse than the result with AdamW.

B. ADAHESSIAN Overhead

Here, we discuss and measure the overhead of ADAHESSIAN. In terms of computational complexity, ADAHESSIAN requires twice the flops as compared to SGD. This 2 \times overhead comes from the cost of computing the Hessian diagonal, when one Hutchinson step is performed per optimization iteration. Each Hutchinson step requires computing one Hessian matvec (the $\mathbf{H}z$ term in Eq. 9). This step requires one more gradient backpropagation, hence leading to twice the theoretical complexity.

We have also measured the actual runtime of ADAHESSIAN in PyTorch on a single RTX Titan GPU machine, as

reported in the second column of Table 8. For ResNet20, ADAHESSIAN is 2.42 \times slower than SGD (and 2.27 \times slower than Adam). As one can see, ADAHESSIAN is not orders of magnitude slower than first order methods. The gap between the measured and theoretical speed is likely due to the fact that Pytorch [45] (and other existing frameworks) are highly optimized for first order methods. Even then, if one considers the fact that SGD needs a lot of tuning, this overhead may not be large.

It is also possible to reduce the ADAHESSIAN overhead. One simple idea is to reduce the Hutchinson calculation frequency from 1 Hessian matvec per iteration to every multiple iterations. For example, for a frequency of 2, we perform the Hutchinson step at every other optimization iteration. This reduces the theoretical computational cost to 1.5 \times from 2 \times . One can also further reduce the frequency to 5, for which this cost reduces to 1.2 \times .

We studied how such reduced Hutchinson calculation frequency approach would impact the performance. We report the results for training ResNet20/ResNet32 on the Cifar10 in Table 8, when we vary the Hutchinson frequency from 1 to 5. As one can see, there is a small performance variation, but the ADAHESSIAN overhead significantly decreases as compared to SGD and Adam.

VI. CONCLUSIONS

In this work, we proposed ADAHESSIAN, an adaptive Hessian based optimizer. ADAHESSIAN incorporates an approximate Hessian diagonal, with spatial averaging and momentum to precondition the gradient vector. This automatically rescales the gradient vector resulting in better descent directions. One of the key novelties in our approach is the incorporation spatial averaging for Hessian diagonal along with an exponential moving average in time. These enable us to smooth noisy local Hessian information which could be highly misleading.

Table 8: Comparison between ADAHESSIAN theoretical and measured speed, as compared to Adam and SGD, tested on Cifar10. We also measured the speed up for different Hessian computation frequencies. As one can see, ADAHESSIAN is not orders of magnitude slower than SGD, despite the widely-held incorrect belief about the efficiency of Hessian based methods. Furthermore, by increasing the Hessian computation frequency, the run time can improve from $3.23\times$ to $1.45\times$, as compared to SGD for ResNet32. The real measurement is performed on one RTX Titan GPU.

Hessian Computation Frequency	1	2	3	4	5
Theoretical Per-iteration Cost (\times SGD)	$2\times$	$1.5\times$	$1.33\times$	$1.25\times$	$1.2\times$
ResNet20 (Cifar10)	$92.13 \pm .08$	$92.40 \pm .04$	$92.06 \pm .18$	$92.17 \pm .21$	$92.16 \pm .12$
Measured Per-iteration Cost (\times SGD)	$2.42\times$	$1.71\times$	$1.47\times$	$1.36\times$	$1.28\times$
Measured Per-iteration Cost (\times Adam)	$2.27\times$	$1.64\times$	$1.42\times$	$1.32\times$	$1.25\times$
ResNet32 (Cifar10)	$93.08 \pm .10$	$92.91 \pm .14$	$92.95 \pm .17$	$92.93 \pm .24$	$93.00 \pm .10$
Measured Per-iteration Cost (\times SGD)	$3.23\times$	$2.12\times$	$1.74\times$	$1.56\times$	$1.45\times$
Measured Per-iteration Cost (\times Adam)	$2.91\times$	$1.96\times$	$1.64\times$	$1.48\times$	$1.38\times$

We extensively tested ADAHESSIAN on various datasets and tasks, using state-of-the-art models. These include IWSLT14 and WMT14 for neural machine translation, PTB and Wikitext-103 for language modeling, GLUE for natural language understanding, Cifar10 and ImageNet for image classification (provided in Appendix IV-B), and Criteo Ad Kaggle for recommendation system (provided in Appendix IV-F). ADAHESSIAN consistently achieves comparable or higher generalization performance as compared to the highly tuned default optimizers used for these different tasks.

Stepping back, it is important for every work to state its limitations (in general, but in particular in this area). The current limitation of ADAHESSIAN is that it is $2 - 3\times$ slower than first order methods such as SGD and Adam. We briefly explored how this overhead could be reduced, but more work is needed in this area. However, ADAHESSIAN consistently achieves comparable or better accuracy. For example, for LM task, ADAHESSIAN achieves up to 2.7 better PPL, as compared to AdamW, which is significant for this task.

Finally, from a higher-level perspective, we should note that there has been significant development within second order methods, both theory and practice, even though these methods were widely viewed as being inapplicable for machine learning even just a few years ago. Some examples include Hessian based model compression [18, 19, 24, 31], adversarial attacks [68], and studies of the loss landscape topology for different NN architectures [52, 66], to name just a few. ADAHESSIAN is an important step in this area, and we expect that it will enable still further progress. We have open sourced ADAHESSIAN and we hope that it would help this progress [1].

ACKNOWLEDGMENTS

This work was supported by a gracious fund from Amazon Machine Learning Research Award (MLRA). The UC Berkeley team also acknowledges gracious support from Intel corporation, Intel VLAB team, Google Cloud, Google TFTC team, and Nvidia. Amir Gholami was supported through through funding from Samsung SAIT. We would like to acknowledge the DARPA, IARPA, NSF, and ONR via its BRC on RandNLA for providing partial support of this work. Our conclusions do not necessarily reflect the position or the policy of our sponsors, and no official endorsement should be inferred.

REFERENCES

- [1] <https://github.com/amirgholami/adahessian.git>, May 2020.
- [2] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.
- [3] Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization in linear time. *Journal of Machine Learning Research*, 1050:15, 2016.
- [4] Amirali Aghazadeh, Vipul Gupta, Alex DeWeese, O Ozan Koyluoglu, and Kannan Ramchandran. Bear: Sketching bfgs algorithm for ultra-high dimensional feature selection in sublinear memory. *arXiv preprint arXiv:2010.13829*, 2020.
- [5] Costas Bekas, Effrosyni Kokiopoulou, and Yousef Saad. An estimator for the diagonal of a matrix. *Applied numerical mathematics*, 57(11-12):1214–1229, 2007.
- [6] Raghu Bollapragada, Richard H Byrd, and Jorge Nocedal. Exact and inexact subsampled newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 2019.
- [7] Raghu Bollapragada, Dheevatsa Mudigere, Jorge Nocedal, Hao-Jun Michael Shi, and Ping Tak Peter Tang. A progressive batching L-BFGS method for machine learning. *arXiv preprint arXiv:1802.05374*, 2018.

- [8] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
- [9] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [10] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyu Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.
- [11] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.
- [12] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.
- [13] Chao Chen, Severin Reiz, Chenhan Yu, Hans-Joachim Bungartz, and George Biros. Fast evaluation and approximation of the gauss-newton hessian matrix for the multilayer perceptron. *arXiv preprint arXiv:1910.12184*, 2019.
- [14] Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. Series on Optimization. SIAM, 2000.
- [15] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformerxl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019.
- [16] Ron S Dembo, Stanley C Eisenstat, and Trond Steihaug. Inexact Newton methods. *SIAM Journal on Numerical analysis*, 19(2):400–408, 1982.
- [17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.
- [18] Zhen Dong, Zhewei Yao, Yaohui Cai, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. HAWQ-V2: Hessian Aware trace-Weighted Quantization of neural networks. *NuerIPS’19 workshop on Beyond First-Order Optimization Methods in Machine Learning*, 2019.
- [19] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 293–302, 2019.
- [20] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
- [21] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.
- [22] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. *arXiv preprint arXiv:1802.09568*, 2018.
- [23] Vipul Gupta, Swanand Kadhe, Thomas Courtade, Michael W Mahoney, and Kannan Ramchandran. Oversketched newton: Fast convex optimization for serverless systems. *arXiv preprint arXiv:1903.08857*, 2019.
- [24] Babak Hassibi and David G Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in neural information processing systems*, pages 164–171, 1993.
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [26] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [27] Forrest N Iandola, Albert E Shaw, Ravi Krishna, and Kurt W Keutzer. Squeezebert: What can computer vision teach nlp about efficient neural networks? *arXiv preprint arXiv:2006.11316*, 2020.
- [28] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.
- [29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.
- [30] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- [31] Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in neural information processing systems*, pages 598–605, 1990.
- [32] Kfir Y Levy. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.
- [33] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.
- [34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [35] Xindian Ma, Peng Zhang, Shuai Zhang, Nan Duan, Yuexian Hou, Ming Zhou, and Dawei Song. A tensorized transformer for language modeling. In *Advances in Neural Information Processing Systems*, pages 2229–2239, 2019.
- [36] James Martens and Roger Grosse. Optimizing neural

- networks with Kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417, 2015.
- [37] H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.
- [38] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.
- [39] Tomáš Mikolov, Anoop Deoras, Stefan Kombrink, Lukáš Burget, and Jan Černocký. Empirical evaluation and combination of advanced language modeling techniques. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [40] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091*, 2019.
- [41] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady an ussr*, volume 269, pages 543–547, 1983.
- [42] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.
- [43] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, 2018.
- [44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- [45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32, pages 8024–8035. Curran Associates, Inc., 2019.
- [47] Jason Phang, Thibault F  vry, and Samuel R Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018.
- [48] Sashank J Reddi, Manzil Zaheer, Suvrit Sra, Barnabas Poczos, Francis Bach, Ruslan Salakhutdinov, and Alexander J Smola. A generic approach for escaping saddle points. *arXiv preprint arXiv:1709.01434*, 2017.
- [49] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- [50] Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled Newton methods I: globally convergent algorithms. *arXiv preprint arXiv:1601.04737*, 2016.
- [51] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.
- [52] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, pages 2483–2493, 2018.
- [53] Tom Schaul, Sixin Zhang, and Yann LeCun. No more pesky learning rates. In *International Conference on Machine Learning*, pages 343–351, 2013.
- [54] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604, 2018.
- [55] Bharat Singh, Soham De, Yangmuzy Zhang, Thomas Goldstein, and Gavin Taylor. Layer-specific adaptive learning rates for deep networks. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 364–368. IEEE, 2015.
- [56] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.
- [57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [58] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
- [59] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, 2019.
- [60] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD’17*, pages 1–7. 2017.
- [61] Shusen Wang, Fred Roosta, Peng Xu, and Michael W Mahoney. Giant: Globally improved approximate newton method for distributed optimization. In *Advances in*

- Neural Information Processing Systems*, pages 2332–2342, 2018.
- [62] Peng Xu, Farbod Roosta-Khorasan, and Michael W Mahoney. Second-order optimization for non-convex machine learning: An empirical study. *arXiv preprint arXiv:1708.07827*, 2017.
 - [63] Peng Xu, Farbod Roosta-Khorasani, and Michael W Mahoney. Newton-type methods for non-convex optimization under inexact Hessian information. *arXiv preprint arXiv:1708.07164*, 2017.
 - [64] Peng Xu, Farbod Roosta-Khorasani, and Michael W. Mahoney. Newton-Type Methods for Non-Convex Optimization Under Inexact Hessian Information. *arXiv preprint arXiv:1708.07164*, 2017.
 - [65] Peng Xu, Jiyan Yang, Farbod Roosta-Khorasani, Christopher Ré, and Michael W Mahoney. Sub-sampled Newton methods with non-uniform sampling. In *Advances in Neural Information Processing Systems*, pages 3000–3008, 2016.
 - [66] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney. Pyhessian: Neural networks through the lens of the hessian. *arXiv preprint arXiv:1912.07145*, 2019.
 - [67] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *Advances in Neural Information Processing Systems*, 2018.
 - [68] Zhewei Yao, Amir Gholami, Peng Xu, Kurt Keutzer, and Michael W Mahoney. Trust region based adversarial attack on neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11350–11359, 2019.
 - [69] Zhewei Yao, Peng Xu, Farbod Roosta-Khorasani, and Michael W Mahoney. Inexact non-convex Newton-type methods. *arXiv preprint arXiv:1802.06925*, 2018.
 - [70] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2019.
 - [71] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.
 - [72] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why adam beats sgd for attention models. *arXiv preprint arXiv:1912.03194*, 2019.
 - [73] Sixin Zhang, Anna E Choromanska, and Yann LeCun. Deep learning with elastic averaging sgd. In *Advances in Neural Information Processing Systems*, pages 685–693, 2015.

APPENDIX

A. Descending Property of Eq. 6

Assume that $f(w)$ is a strongly convex and strictly smooth function in \mathbb{R}^d , such that there exists positive constants α and β so that $\alpha I \leq \nabla^2 f(w) \leq \beta I$ for all w . We can show that the update formulation of Eq. 6 is a converging algorithm. Particularly, we can show that with the proper learning rate:

$$f(w_{t+1}) - f(w_t) \leq -\frac{\alpha^k}{2\beta^{1+k}} \|\mathbf{g}_t\|^2. \quad (13)$$

Note that when $k = 0$ or 1 , the convergence rate is the same as gradient descent or Newton method⁶ [9], respectively. Our proof is similar to [9] for Newton method.

Let us define $\lambda(w_t) = (\mathbf{g}_t^T \mathbf{H}_t^{-k} \mathbf{g}_t)^{1/2}$. Since $f(w)$ is strongly convex, we have

$$\begin{aligned} f(w_t - \eta \Delta w_t) &\leq f(w_t) - \eta \mathbf{g}_t^T \Delta w_t + \frac{\eta^2 \beta \|\Delta w_t\|^2}{2} \\ &\leq f(w_t) - \eta \lambda(w_t)^2 + \frac{\beta}{2\alpha^k} \eta^2 \lambda(w_t)^2. \end{aligned} \quad (14)$$

The last inequality comes from the fact that

$$\lambda(w_t)^2 = \Delta w_t^T \mathbf{H}_t^k \Delta w_t \geq \alpha^k \|\Delta w_t\|^2. \quad (15)$$

Therefore, the step size $\hat{\eta} = \frac{\alpha^k}{\beta}$ will make f decreases as follows,

$$f(w_t - \hat{\eta} \Delta w_t) \leq f(w_t) - \frac{1}{2} \hat{\eta} \lambda(w_t)^2. \quad (16)$$

Since $\alpha \preceq \mathbf{H}_t \preceq \beta$, we have

$$\lambda(w_t)^2 = \mathbf{g}_t^T \mathbf{H}_t^{-k} \mathbf{g}_t \geq \frac{1}{\beta^k} \|\mathbf{g}_t\|^2. \quad (17)$$

Therefore,

$$f(w_t - \hat{\eta} \Delta w_t) - f(w_t) \leq -\frac{1}{2\beta^k} \hat{\eta} \|\mathbf{g}_t\|^2 = -\frac{\alpha^k}{2\beta^{1+k}} \|\mathbf{g}_t\|^2.$$

B. Descending Property of Eq. 7

When $f(w)$ is a strongly convex and strictly smooth function in \mathbb{R}^d , such that there exists positive constants α and β so that $\alpha I \leq \nabla^2 f(w) \leq \beta I$ for all w , we can prove that Eq. 7 has the same convergence rate as Eq. 6.

First of all, it is not hard to see the diagonal elements in \mathbf{D} are all positive since $f(w)$ is a strongly convex problem. That is,

$$\alpha \leq e_i^T \mathbf{H} e_i = e_i^T \mathbf{D} e_i = D_{i,i}, \quad (18)$$

where e_i is the vector whose coordinates are all zero, except the i -th one that equals 1. Similarly, we have

$$D_{i,i} = e_i^T \mathbf{D} e_i = e_i^T \mathbf{H} e_i \leq \beta. \quad (19)$$

Therefore, the diagonal elements in \mathbf{D} are in the range $[\alpha, \beta]$. Using the same proof as in Appendix A, we will get the result.

C. Descending Property of Eq. 10

When $f(w)$ is a strongly convex and strictly smooth function in \mathbb{R}^d , such that there exists positive constants α and β so that $\alpha I \leq \nabla^2 f(w) \leq \beta I$ for all w , we can prove that Eq. 10 has the same convergence rate as Eq. 6.

As shown in Appendix B, the diagonal elements in \mathbf{D} are in the range $[\alpha, \beta]$. Therefore, the average of a subset of those numbers is still in the range $[\alpha, \beta]$. Using the same proof as in Appendix A, we will get the result.

⁶The convergence rate here denotes the global convergence rate of Newton's method.

D. Experimental Setup

Here, we provide more details on the experimental setup for the empirical evaluation.

a) *Image Classification*: The training/test sets for Cifar10 [30] dataset contain 50k/10k images, respectively. The models used on Cifar10 are standard ResNet20/32. We train both models with 160 epochs and decay the learning rate by a factor of 10 at epoch 80 and 120. The batch size is set to be 256. For SGD/Adam/AdamW, the initial learning rates are tuned and set to be 0.1/0.001/0.01. For ADAHESSIAN, we set the block size as 9, k to be 1, and learning rate as 0.15 for both ResNet20/32. For Adam/AdamW/ADAHESSIAN, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We run each experiment 5 times on Cifar10 and report the mean and standard deviation of the results. The training/test sets for ImageNet dataset [17] contain 1.2M/50k images, respectively. Our code is modified from the official PyTorch example⁷. The batch size is set to be 256. We train ResNet18 for 90 epochs. All the settings of different optimizers are the same as used in Cifar10 example.

b) *Neural Machine Translation*: The training/validation/test sets for the IWSLT14 dataset contain about 153K/7K/7K sentence pairs, respectively. We use a vocabulary of 10K tokens via a joint source and target byte pair encoding (BPE). For the WMT14 dataset, we follow the setting of [57], which contains 4.5M parallel sentence pairs for training. We use Newstest2014 as the test set, and Newstest2013 as the validation set. The 37K vocabulary for WMT14 is also via a joint source and target BPE factorization. We set dropout as 0.0 for Transformer base/small model. For AdamW, we follow the optimizer setting and learning rate schedule in [59]. For ADAHESSIAN, we set the block size as 32 for IWSLT/WMT, k to be 1.0, and learning rate as 0.047/1.0 for IWSLT/WMT. For both AdamW/ADAHESSIAN, we set $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We fix the label smoothing value as $\epsilon_{ls} = 0.1$ in all experiments. We implement our code for MT based on *fairseq-py* [42]. We employ BLEU⁸ [44] as the evaluation metric for MT. Following standard practice, we measure tokenized case-sensitive BLEU and case-insensitive BLEU for WMT14 En-De and IWSLT14 De-En, respectively. For a fair comparison, we do not include other external datasets. We train 130/55 epochs for WMT/IWSLT, respectively. We set the maximum token size to be 4096×8 (eight gpus)/4096 (one gpu) for WMT/IWSLT. For inference, we average the last 10/5 checkpoints, and we set the length penalty as 0.6/1.0 and beam size as 4/5 for WMT/IWSLT, following [42]. We run each experiment 5 times on IWSLT and report the mean and standard deviation of the results.

c) *Language Modeling*: PTB [39] has 0.93M training tokens, 0.073M validation tokens, and 0.082M test tokens. Wikitext-103 [38] contains 0.27M unique words, and 100M training words from 28K articles, with an average length of 3.6K words per article. We use the same evaluation scheme following [15]. We use a three-layer tensorized transformer core-1 for PTB and a six-layer tensorized transformer core-1 for Wikitext-103, following [35]. We set the dropout rate as 0.3 in all the LM experiments. The model is trained for 30 epochs on both PTB and WikiText-103. For AdamW, we follow the learning rate setting in [35]. For ADAHESSIAN, we set the block size as 4 and k as 0.5 for PTB and Wikitext-103. We set the learning rate as 2.0/1.0 for PTB/Wikitext-103, respectively. We set the batch size to be 60/120 for PTB/Wikitext-103. For AdamW/ADAHESSIAN, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the warmup steps to be 4000 and label smoothing to be $\epsilon_{ls} = 0.1$ in all LM experiments. We run each experiment 5 times on PTB and report the mean and standard deviation of the results.

d) *Natural Language Understanding*: The SqueezeBERT model is pre-trained as the authors suggested [27]. Particularly, we pretrain SqueezeBERT from scratch using the LAMB [70] optimizer with a global batch size of 8192, a learning rate of $2.5e-3$, and a warmup proportion of 0.28. We pretrain for 56k steps with a maximum sequence length of 128 and then for 6k steps with a maximum sequence length of 512 followed [26]. Moreover, before directly using the pre-trained model on GLUE tasks, we apply transfer learning from the MNLI GLUE task [58] to other GLUE tasks [26, 47]. We refer readers to [26] for more detailed instruction.

For finetuning, we set the batch size as 16, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ for both AdamW/ADAHESSIAN as suggested in [26]. For ADAHESSIAN, we set the block size b as 4 and k as 1 for all tasks. As in [26] we perform hyperparameter tuning on the learning rate and dropout rate.

e) *Recommendation System*: The Criteo Ad Kaggle dataset consists of click logs for ad CTR prediction for 7 days. Each data set contains 13 continuous and 26 categorical features. The dataset contains about 45 million

⁷<https://github.com/pytorch/examples/tree/master/imagenet>

⁸<https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl>

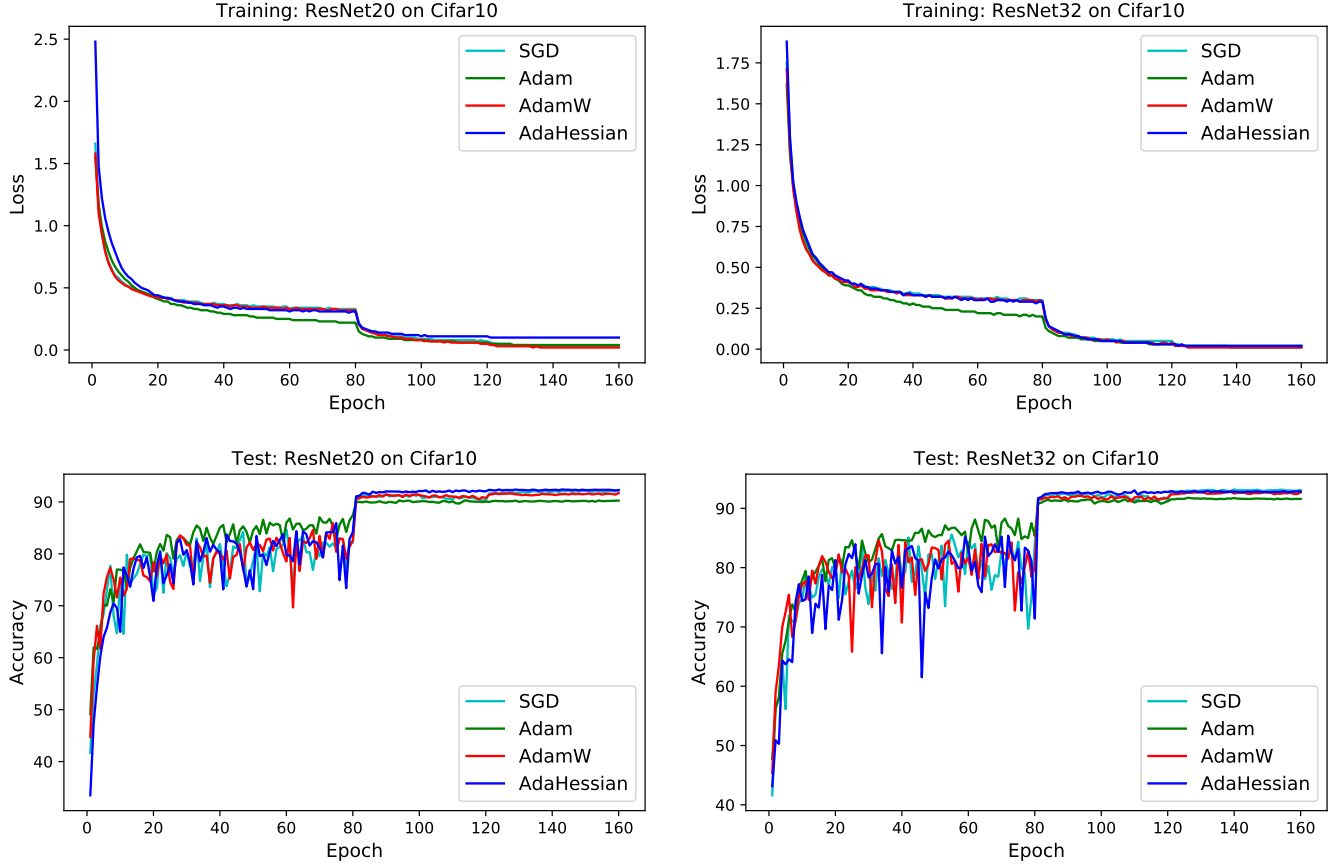


Fig. A.7: Training and testing loss curves of SGD, Adam, AdamW, ADAHESSIAN for ResNet20/32 on Cifar10. SGD and ADAHESSIAN consistently achieve better accuracy as compared to Adam and AdamW. The final accuracy results are reported in Table 2.

samples over 7 days. In experiments, we follow the setting from [40]. Our code is also modified from [40]⁹. The testing metric for Recommendation Systems is Click Through Rate (CTR), measured on training and test sets. For Adagrad, the learning rate is set to be 0.01. For ADAHESSIAN, we set the block size as 1, k as 0.5, learning rate as 0.043, $\beta_1 = 0.9$, and $\beta_2 = 0.98$. We set the batch size to be 128, following [40].

f) Delayed Hessian Update: For ResNets on Cifar10, we use 5 epochs for warmup. In particular, within 5 epochs, the Hessian diagonal is still computed for every iteration. After that, the Hessian diagonal computation frequency is set to be between 1 to 5 iterations.

E. Additional Results

In this section, we present additional empirical results that were discussed in Section IV. See Figure A.7 and A.8.

⁹<https://github.com/facebookresearch/dlrm>

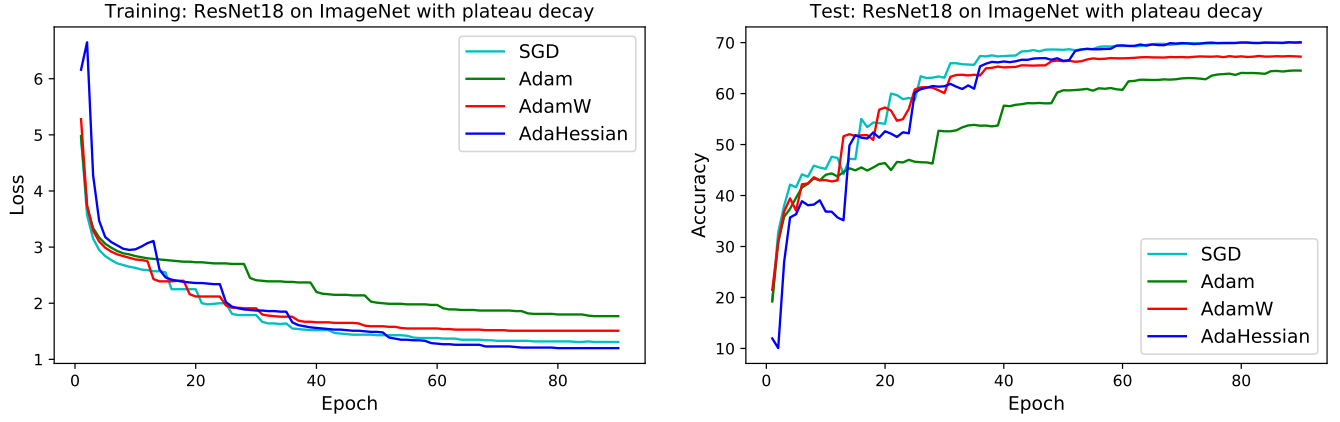


Fig. A.8: Training/Test loss curve of SGD, Adam, AdamW, ADAHESSIAN for ResNet18 on ImageNet. SGD and ADAHESSIAN consistently achieve better accuracy as compared to Adam and AdamW. The final accuracy results are reported in Table 2.

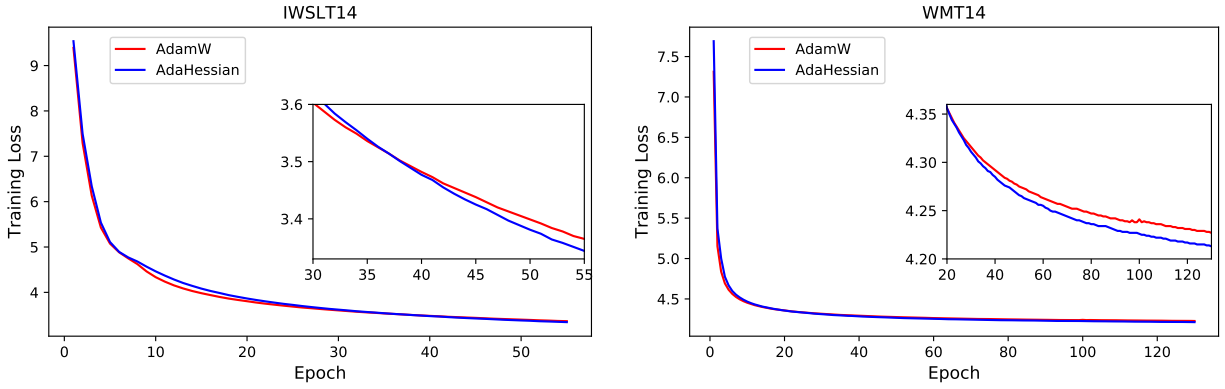


Fig. A.9: Training loss curves of AdamW and ADAHESSIAN for Transformer on IWSLT14 and WMT14. The training loss of ADAHESSIAN is lower than that of AdamW on both IWSLT14 and WMT14. Testing results are reported in Table 3.

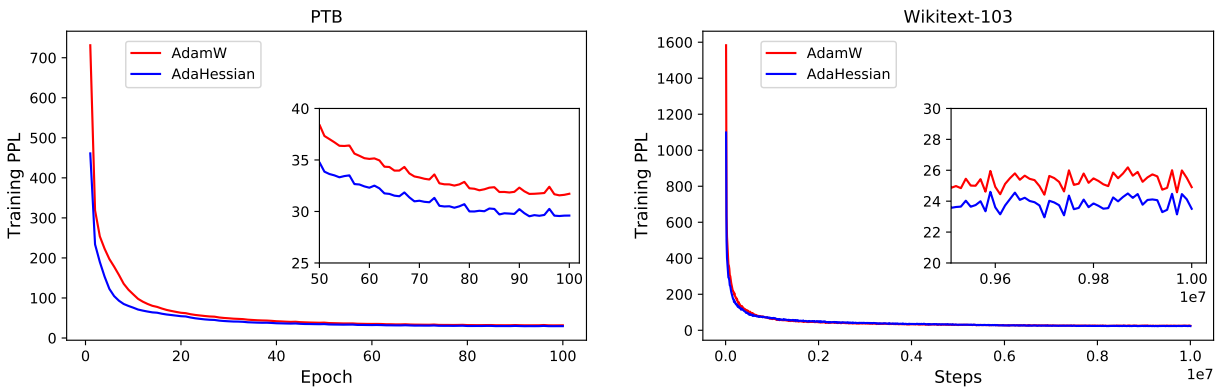


Fig. A.10: Training PPL curves of AdamW and ADAHESSIAN for Transformer on PTB and Wikitext-103. The losses of ADAHESSIAN are consistently lower than AdamW from the beginning of the training. ADAHESSIAN achieves 29.56/23.51 final training perplexity (PPL) on PTB/Wikitext-103 as compared to AdamW (31.72/24.01). Testing results are reported in Table 4.