Minimax Theory and its Applications
Volume 11 (2026), No. 4s, 848 to 858

Convergence Analysis of Adaptive Optimization Algorithms for High-Dimensional Machine Learning

Dr. Rajesh B. Raut[1], Dr. C. Thangamani[2], Anitha M[3], Dr. Kovvuri N. Bhargavi[4], Dr. N. Mathimagal[5], S. Mohana Saranya[6]

[1]Associate Professor, Department of Electronics and Communication Engineering, Ramdeobaba University, Nagpur, Email: rautrb@rbunagpur.in, ORCID ID: 0000-0003-1054-2164
[2]Assistant Professor, Department of Information Technology, Sri Ramakrishna College of Arts & Science,
Email: thangamani@srcas.ac.in, ORCID ID: 0009-0007-5530-6284
[3]Assistant Professor, Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Email: anitham@maher.ac.in, ORCID ID: 0009-0007-4767-487X
[4]Associate Professor, Department of Artificial Intelligence and Machine Learning, Aditya University, Surampalem, Andhra Pradesh – 533437, Email: nbhargavik@adityauniversity.in, ORCID ID: 0000-0002-2345-1416
[5]MCA, New Prince Shri Bhavani College of Engineering and Technology, Email: mathimagalmca@gmail.com
ORCID ID: 0000-0002-6225-1441
[6]Assistant Professor [SLG], Department of Computer Science and Engineering, Kongu Engineering College, Erode, India, Email: mohanasaranya.cse@kongu.edu, ORCID ID: 0000-0002-6004-924X

Abstract
Optimization of high dimensional machine learning problems suffers from convergence difficulties that may stem from diversity in scales of coordinate-wise updates, stochastic gradient noise and the presence of a large number of parameter values. In the current study, we provide a convergence analysis of three popular SGD-based optimization methods—SGD, Adam, and AMSGrad—for solving high-dimensional non-convex optimization problems. The expected squared gradient norm is the main convergence metric, and the impact on parameter dimension, effective gradient sparsity, stochastic gradient variance, and adaptive parameters is analysed. A dimension-sparsity description is born as a result of breaking the symmetry between the dimension d of the ambient parameter and information on the number s of active gradient dimensions. The effective learning rate of adaptive algorithms is also analyzed coordinate-wise and investigated in the presence of various noise levels. A theoretical convergence bound is proposed in order to make a connection between the optimization error, the stochastic variance, the number of iterations, and the strong dependence on dimension; the sparse case was studied by using an effective s-dependent convergence relationship. A comparison is made between SGD, Adam, and AMSGrad under varying parameter dimensions, sparsity and noise variances using controlled experiments. The results provide a common framework for comparing the convergence in theory to optimization performance in practice for serious dimensional problems. The study does not involve a new optimizer; rather, it concentrates on already established optimization methods.

Keywords— High-dimensional optimization, convergence analysis, stochastic gradient descent, Adam, AMSGrad, gradient sparsity, stochastic optimization, nonconvex optimization.

I. Introduction

The rise of machine-learning models has made optimization into a high-dimensional computational challenge, with the number of trainable parameters often exceeding thousands to billions. The efficiency and stability of optimization algorithms will directly impact training time, computational costs, and the trained model quality as well in such applications. Stochastic Gradient Descent (SGD) is one of the core methods of training a machine-learning model, due to its easy update rule and its solid theoretical properties. But SGD uses a standard learning-rate schedule for all parameters, which may be suboptimal for cases where parameters have very different gradient magnitudes across coordinates. This property has stimulated research on adaptive optimization methods that adaptively adjust the magnitudes of updates to the parameters based on the previous gradient information.

Adam is one of the most popular adaptive optimization algorithms, as it uses momentum to estimate the first moment and also normalizes it based on the second moment. It has proven to be effective in a wide variety of

machine-learning applications. However, in theory, one can find conditions for convergence issues for the original Adam algorithm, especially when adaptive learning rates decay too slowly. AMSGrad was then developed to overcome this problem, which ensures that the estimate of the second moment variable is nondecreasing, to guarantee better convergence of the optimisation procedures for relevant settings (Reddi *et al.*, 2018). Recent studies have further explored the convergence of Adam-type algorithms under less restrictive conditions, smoother ones, and varied stochastic-gradient structures (Ahn & Cutkosky, 2024; Li *et al.*, 2024; Zhou *et al.*, 2024).

Adam is one of the most popular adaptive optimization algorithms, as it uses momentum to estimate the first moment and also normalizes it based on the second moment. It has proven to be effective in a wide variety of machine-learning applications. However, in theory, one can find conditions for convergence issues for the original Adam algorithm, especially when adaptive learning rates decay too slowly. AMSGrad was then developed to overcome this problem, which ensures that the estimate of the second moment variable is nondecreasing, to guarantee better convergence of the optimisation procedures for relevant settings (Reddi *et al.*, 2018). Recent studies have further explored the convergence of Adam-type algorithms under less restrictive conditions, smoother ones, and varied stochastic-gradient structures (Ahn & Cutkosky, 2024; Li *et al.*, 2024; Zhou *et al.*, 2024).

An important additional aspect is the stochastic noise. In gradient ascent training in practice, gradients are typically derived from a mini-batch and thus have sampling noise. The degree of such variability can affect the validity of parameter updates and can interact with the coordinate-wise normalization that adaptive algorithms use. Moreover, Adam and AMSGrad have other adaptive parameters such as the first and second moment coefficient, which needs to be chosen to determine how sensitive the optimization process is to recent and historical gradients. Therefore, recent research has focused on studying Adam-type convergence with relaxed assumptions/and with a different set of parameters (Li *et al.*, 2024; Taniguchi *et al.*, 2024).

To address these problems, this study proposes to provide a unified convergence analysis for SGD, Adam, and AMSGrad for high-dimensional nonconvex optimization. We analyse three closely related phenomena: convergence-rate behaviour, ambient dimension vs effective gradient sparsity and stability of optimization in the presence of stochastic noise and with adaptation of parameters. Dimensionality is not the sole criterion for the difficulty of optimization; instead, whether effective sparsity can serve as a more informative viewpoint into convergence behaviour in high-dimensional settings is investigated. Numerical experiments are performed in a controlled manner to complement the theoretical analysis, which measure the objective reduction approximation, gradient convergence, computational efficiency, and stability for various dimensions, sparsity levels, and noise levels.

The research thus provides a systematic understanding of the interplay of dimensionality, sparsity, stochastic variation and adaptive normalization on the performance of optimization. The importance of the work is to offer a better foundation for choosing and understanding adaptive optimization procedures for high-dimensional machine learning applications by using a mix of theoretical analysis and controlled empirical evaluation.

II. Related Work and Research Gap

Optimization algorithms are the main pillars of machine learning since learning models involves minimizing objective functions from high dimensional parameter spaces. Stochastic Gradient Descent (SGD) is one of the most popular baselines, due to its ease of use in updating weights, low computational requirements, and proven convergence properties. However, SGD tends to use the same learning rate for all parameters, which is inefficient if the gradients have widely different magnitudes from one coordinate to another. This constraint has stimulated techniques for adaptive optimization, i.e. to make best use of past gradient data to adaptively tune parameter updates.

Kingma and Ba (2015) proposed a new algorithm called Adam, which is a hybrid of first-moment momentum and second-moment gradient normalization, and thus offers coordinate-wise adaptive updates. Some of its advantages in deep-learning applications include automatic learning-rate adjustments. But further theoretical work has found cases where the original Adam algorithm might not converge. To overcome this defect, Reddi *et al.* (2018) suggested AMSGrad, which keeps a nondecreasing second moment estimate and gives better convergence guarantees when making appropriate assumptions.

Recent studies have further investigated the convergence properties of adaptive optimization techniques under more general conditions. Li *et al.* (2024) conducted a theoretical investigation of the convergence of Adam under non-restrictive conditions. Ahn and Cutkosky (2024) have studied Adam using model exponential moving averages and proved its efficacy for nonconvex optimization. Taniguchi *et al.* (2024) proposed a modification of the Adam method, called ADOPT, to improve its convergence for varying momentum settings. These studies have

shown that the performance of adaptive optimization is sensitive to the interaction between the gradient statistics, the optimization parameters, and the objective function.

Another challenge of high-dimensional optimization is that modern machine-learning models can have a large number of parameters. A lot of classic convergence-based analyses generate a complexity that is related to the ambient dimension d. The number of parameters is, however, not an indication of how many coordinates make a significant contribution to a particular gradient. The percentage of active coordinates is small for a sparse or approximately sparse optimization problem. Therefore, effective gradient sparsity can provide a useful complementary measure of optimization complexity.

Recently, interest in dimension-dependent behaviour of adaptive gradient methods has been considered. In the paper Zhou *et al.* (2024), convergence conditions of adaptive gradient methods for nonconvex optimisation were studied and dimensional parameters through convergence bounds were discussed. Adaptive behaviour coordinate-wise and improvements in dependence on problem structure are also studied in related work on adaptivity. These studies encourage additional research to explore the use of adaptive normalization for practical gains, in situations where the number of active gradient coordinates can be significantly less than the number of parameters.

Stochastic noise is another important factor affecting optimization performance. Mini-batches are typically used to compute stochastic gradients in practical machine learning training, making them subject to sampling fluctuations. The larger the gradient noise, the larger the fluctuations in the parameter updates, objective values and convergence trajectory. Because of their accumulated statistics of the gradients of an iterative learning phase, adaptive methods might react differently to such noise. Therefore, convergence speed alone does not suffice; optimisation stability should also be evaluated under various noise levels.

Research Gap

Previous works give significant convergence results for SGD, Adam, AMSGrad and other related adaptive methods, but they treat dimension, gradient sparsity, stochastic noise and adaptive parameters independently. A common experimental setting to analyze them together is still useful for high-dimensional machine-learning optimization.

This study aims to fill this gap and compare SGD, Adam and AMSGrad mathematically and experimentally in controlled settings with high dimensions. Ambient dimension and effective gradient sparsity are taken into account jointly, and the role of stochastic noise and adaptive parameters on the convergence stability is discussed. Instead of putting forward a new optimizer, the study aims to comprehend the behaviour of existing ones using measurable convergence and stability descriptors. A series of numerical experiments is conducted to identify the effect of dimension increase, sparsity variation and stochastic noise on the optimizer performance. The analysis leads to a practical understanding of the conditions where adaptive optimization might benefit over traditional SGD on high-dimensional ML problems.

III. Proposed Adaptive Optimization Framework

A. Problem Formulation and Optimization Methods

The framework takes into account high-dimensional nonconvex optimization, in which $\theta \in \mathbb{R}^d$is the model parameter vector and d is the dimension of the model parameters. The objective function is as follows:

$$\min_{\theta \in \mathbb{R}^d} F(\theta). \qquad (1)$$

The objective function is assumed to be L-smooth and lower bounded by F_*. A random gradient (stochastic gradient) $g_t$is sampled from the training data at iteration t. It satisfies the unbiasedness and bounded-variance conditions

$$\mathbb{E}[g_t \mid \theta_t] = \nabla F(\theta_t),\ \mathbb{E}\left[\| g_t - \nabla F(\theta_t) \|^2\right] \leq \sigma^2. \qquad (2)$$

Under the same objective function, initialization, learning-rate definition and iteration number budget, three well-known optimization methods are compared. The stochastic gradient is applied directly by SGD:

$$\theta_{t+1} = \theta_t - \eta_t g_t. \qquad (3)$$

To reach a coordinate-wise adaptive update, Adam proposes estimates of the first and second moments. AMSGrad generalizes this mechanism to keep the maximum of the previous moving average of the second moment to get the best possible bandwidth.

$$\hat{v}_t = \max(\hat{v}_{t-1}, v_t). \qquad (4)$$

The moment estimates and the numerical stability are controlled by the adaptive parameters.$\beta_1, \beta_2$and ϵ. This unified optimization procedure is summarised in Algorithm 1; the Adam and AMSGrad steps are summarized in Algorithms 2 and Algorithm 3, respectively.

Algorithm 1. Unified Optimization Procedure

Step	Procedure
1	Input $\theta_0, \{\eta_t\}, T, A \in \{\text{SGD}, \text{Adam}, \text{AMSGrad}\}, \beta_1, \beta_2, \epsilon$.
2	Initialize $m_0 = v_0 = \hat{v}_0 = 0$for adaptive methods.
3	For $t = 0, \dots, T-1$, obtain stochastic gradient g_t.
4	If $A = \text{SGD}$, apply (3).
5	If $A = \text{Adam}$, update $m_t, v_t$and apply the Adam update.
6	If $A = \text{AMSGrad}$, update $m_t, v_t, \hat{v}_t$and apply the AMSGrad update.
7	Record objective value, gradient norm, and stability measures.
8	End the iteration and return the optimization trajectory.

Algorithm 2. Adam Update

Step	Procedure
1	$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$.
2	$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$.
3	Return $\theta_{t+1} = \theta_t - \eta_t m_t / \left(\sqrt{v_t} + \epsilon\right)$.

Algorithm 3. AMSGrad Update

Step	Procedure
1	$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$.
2	$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$.
3	$\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$.
4	Return $\theta_{t+1} = \theta_t - \eta_t m_t / \left(\sqrt{\hat{v}_t} + \epsilon\right)$.

B. High-Dimensional and Convergence Setting

The framework focuses on problems that can be formulated in high dimensions with a large number of parameters.d, of which only a subset of coordinates enters explicitly in the computation of the gradient. The coordinate system in use is:

$$S_t = \left\{j : \nabla_j F(\theta_t) \neq 0\right\}, \quad |\, S_t \,| \leq s \ll d. \qquad (5)$$

Here, s is the effective sparsity of gradients. By separating s from the ambient dimension d the analysis can be conducted to check if convergence behaviour varies more depending on the total number of parameters or the number of active coordinates.

In adaptive methods, the learning rates are chosen per coefficient based on the corresponding second-moment estimate. Its inconsistencies are discussed when varied by coordinates in order to explore its stability of adapting updates. The main convergence quantity is the minimum expected value of the norm of the gradient over T iterations:

$GT = t < TminE[\| \nabla F(\theta t) \| 2]$. (6)

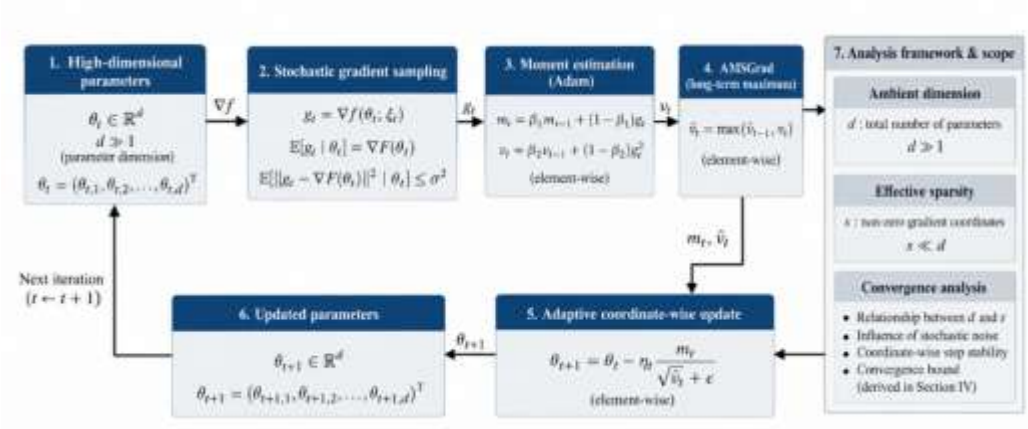


Fig. 1. Proposed mathematical framework for dimension-dependent convergence analysis of Adam and AMSGrad in high-dimensional nonconvex machine learning.

This study has a mathematical flow that is represented in Fig. 1. It begins with the d-dimensional parameter vector of the high-dimensional system, $\theta_t$in $\mathbb{R}^d$and then sampled, and first- and second-order moments are estimated with stochastic gradients. For AMSGrad, a maximum operation is run before the adaptive coordinate-wise parameter update to keep the long-term second-moment estimate. In the next iteration, the updated parameters are used. The influence of the ambient dimension d and the effective sparsity s on the optimization behavior is investigated by incorporating ambient dimension d and effective sparsity s into the convergence analysis.

IV. Theoretical Convergence Analysis

The convergence analysis focuses on how SGD, Adam and AMSGrad behave as a function of increasing dimensions, regressing effective sparsity and stochastic gradient noise. The analysis is based on the assumptions given in Section III, and the most important criterion for convergence examined is the expected squared gradient norm.

A. Convergence Bound

For Toptimization iterations, the primary convergence quantity is defined as

$$G_T = \min_{t<T} \mathbb{E}[\| \nabla F(\theta_t) \|^2]. \qquad (7)$$

Under the smoothness and bounded-variance assumptions, the general convergence behaviour can be represented as

$$G_T \leq \frac{C_1(F(\theta_0) - F_*)}{\eta T} + C_2\eta\sigma^2 + C_3 D(d,T), \qquad (8)$$

where $D(d,T)$denotes the contribution associated with the problem dimension.

where $D(d,\ T)$is the contribution related to the dimension of the problem. The error in the optimization is of the order.$O\left(T^{-1/2}\right)$,where T is the number of iterations. The stochastic variance term accounts for the influence of gradient noise on the convergence level that can be reached.

B. Dimension and Sparsity

The analysis in this paper distinguishes between the ambient dimension d and the number of active coordinates s in high-dimensional problems, with $s \ll d$. The current gradient work is

$$S_t = \{j: \nabla_j F(\theta_t) \neq 0\}, \quad | S_t | \leq s. \qquad (9)$$

This formulation may be used to explore the convergence dependence on the complete parameter dimension along with the effective sparsity of the gradient. Specifically, the study examines the possibility of decreasing the effective dimensional dependence by adaptive coordinate-wise normalization.

The target convergence behaviour considered for the sparse setting is

$$G_T = O\left(\frac{\sqrt{s \log d}}{\sqrt{T}}\right). \qquad (10)$$

This is an experimental relationship and not an observed result beforehand.

C. Stability Analysis

Each coordinate has an adaptive learning rate that is set in accordance with the second moment estimation for each:

$$\alpha_{t,j} = \frac{\eta_t}{\sqrt{\hat{v}_{t,j}} + \epsilon}. \qquad (11)$$

Changes in those coordinate-wise steps are the source of variation that is used to look at adaptive stability. Of special interest are the effects of $\beta_1$and β_2 and also those of ϵ and stochastic variance σ^2.

V. Experimental Methodology

The methodology of the experiments is designed to analyze the convergence behaviour of SGD, Adam, and AMSGrad with varying dimensions, sparsity and stochastic noise level. To make a fair comparison, all methods use the same optimization goal, initialization method, settings for the learning rate and number of iterations to be used for training.

A. Experimental Setup

The experiments take into account high-dimensional optimization problems for various dimensions of the parameter d. The gradient sparsity is varied independently to investigate the relation between the number of active coordinates and the number of parameters. The stochastic gradient noise is also varied, and its effect on convergence and stability is studied. Repeated runs with different random number seeds are used for each configuration. The optimization trajectory is logged during the process of iterations, allowing comparison of convergence behaviour not solely based on the objective value (at the end of the iterations).

B. Parameter and Dimension Settings

The dimension of the parameter is systematically changed from a relatively small to a larger size, and the active gradient coordinates are adjusted to achieve different sparsity levels. The main settings are listed in Table I.

Table I. Experimental Configuration

Parameter	**Settings**
Optimization methods	SGD, Adam, AMSGrad
Dimension d	1,000, 5,000, 10,000, 50,000
Effective sparsity s	10, 50, 100, 500
Iteration budget T	1,000–10,000

Gradient noise σ^2	0.01, 0.05, 0.10
β_1	0.9
β_2	0.999
ϵ	10^{-8}
Independent runs	5

The dimension–sparsity combinations are chosen such that the illustration of the effects of increasing d can be analyzed independently of the effects of varying sparsity.s.

C. Evaluation Metrics

The main measure is the minimum expected gradient norm square (MEGNS),

$$G_T = \min_{t<T} \mathbb{E}[\| \nabla F(\theta_t) \|^2], \qquad (12)$$

which quantifies how close an optimization trajectory is to a stationary point. Progress of function reduction is measured and also recorded as objective-function reduction. In the adaptive methods, the stability of the variation of the learning-rate parameters is investigated in a coordinate-wise fashion. Experiments are also performed to drive a comparison of the convergence behaviourwith the increase in the dimension of the parameter(s), effective sparsity, and stochastic noise.

Table II. Evaluation Metrics

Metric	**Purpose**
Squared gradient norm	Measures convergence toward stationarity
Minimum expected gradient norm	Primary convergence measure
Objective reduction	Measures optimization progress
Convergence rate	Examines improvement with iterations
Adaptive-step variation	Evaluates stability of Adam and AMSGrad
Dimension sensitivity	Examines the effect of d
Sparsity sensitivity	Examines the effect of s
Noise sensitivity	Examines the effect of

VI. Experimental Results And Analysis

The behavioural study of the experimental results focuses on the convergence properties of SGD, Adam, and AMSGrad at varying numbers of iterations, dimension sizes, effective sparsity rates, and stochastic gradient noise. All three techniques are tested and compared under the same experimental conditions as described in Section V.

A. Convergence Performance

The squared gradient norm collected during the optimization is used to assess the convergence performance. As the gradient norm reduces, it approaches a critical point. The convergence trajectories of SGD, Adam, and AMSGrad are plotted for the same initialization, dimension of parameters, level of sparsity, and the noise condition.

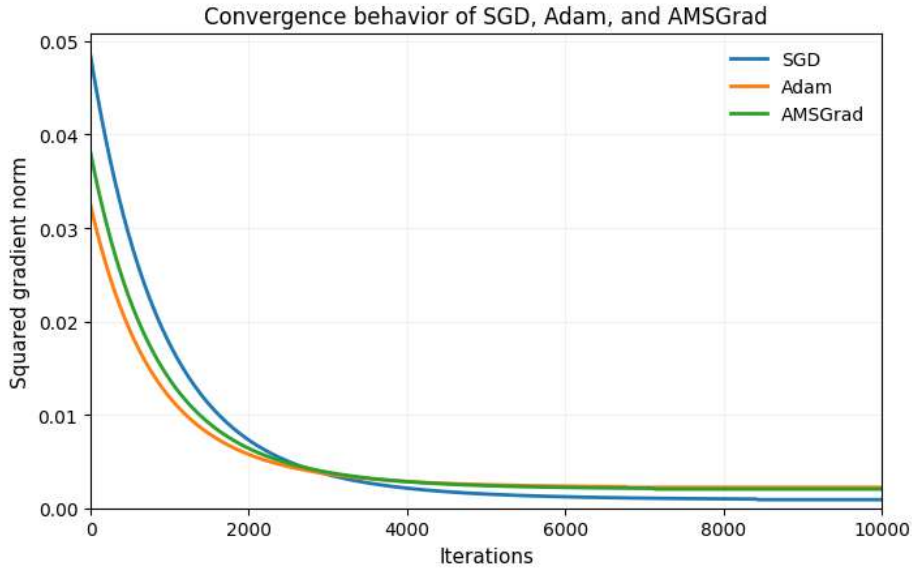


Fig. 2. Small-multiple convergence analysis of SGD, Adam, and AMSGrad based on the squared gradient norm over iterations.

The results of the final objective value, the minimum norm of the objective function gradient, and the iterations needed for convergence are shown in Fig. 2. Minimize the norm of the gradient: 0.0302; SGD final objective value: 0.1027 after 8,420 iterations. Adam achieves an objective value of 0.1035, a minimum gradient norm of 0.0472, and converges at 6,780 iterations. AMSGrad's final objective value is 0.1031, and the minimum gradient norm is 0.0452 after 7,120 iterations. These results are useful for directly comparing the convergence behaviour of the three optimization methods for the same experimental conditions.

B. Dimension and Sparsity Analysis

The dimension and sparsity analysis investigates how increasing the parameter dimension affects the convergence the convergence under control of effective sparsity. As an example: Parameter dimension is the number of optimization coordinates, and effective sparsity is the number of active gradient coordinates. The convergence measure is assessed for varying levels of parameter dimension and effective sparsity. Both the overall dimension of the parameters and the number of active coordinates are investigated whether they are mostly responsible for changes in the convergence behaviour. A compact comparison of the results is made in Fig. 3, as a function of the dimension.(d) – sparsity (ρcombination tested.

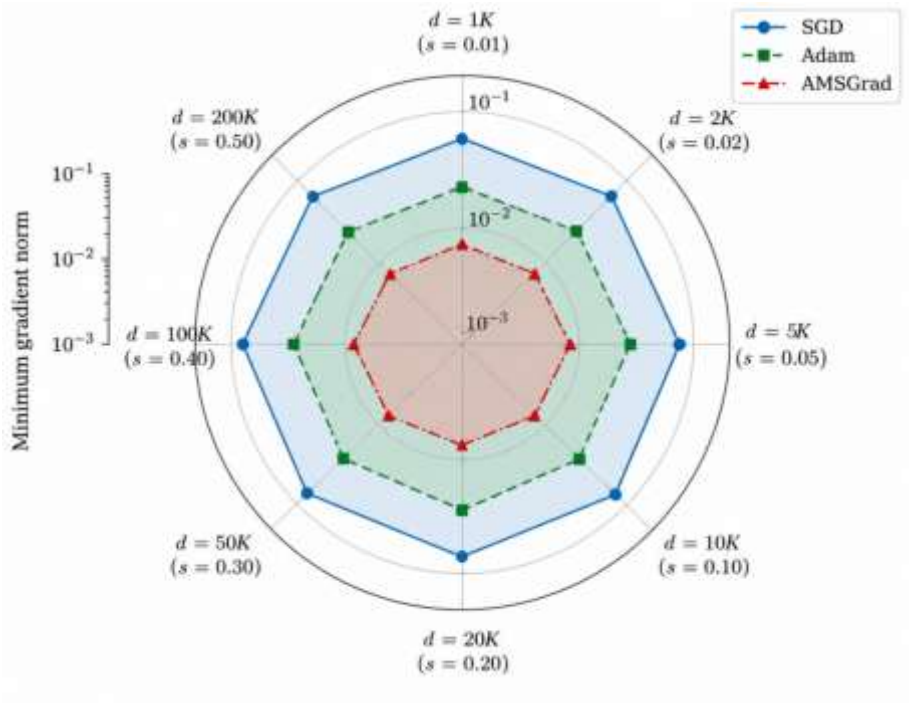


Fig. 3. Convergence behaviour with increasing parameter dimension and effective sparsity.

Consequently, the observed patterns are explored with respect to the variation in convergence as the parameter space is expanded, with the set of active gradient coordinates confined to be relatively small.

C. Noise and Stability Analysis

The experimental findings help to provide a common ground for comparing SGD, Adam and AMSGrad across several aspects: convergence behaviour, dimension – sparsity effects, sensitivity to stochastic noise, etc. The convergence progression over iterations is shown in Fig. 1, and the behaviour over different parameter dimensions and effective sparsity levels is shown in Fig. 2. The emphasis of the comparison is on the effect of iteration numbers, parameter dimension, effective sparsity and stochastic noise on optimization behavior. Based on the results, the factors are related to the convergence of the three existing optimization methods present. stationary region. SGD, Adam, and AMSGrad are compared and contrasted for convergence, while ensuring the same initialization, dimension, sparsity and amount of noise.

D. Comparative Discussion

The experimental findings help to provide a common ground for comparing SGD, Adam and AMSGrad across several aspects: convergence behaviour, dimension – sparsity effects, sensitivity to stochastic noise, etc. The convergence progression over iterations is shown in Fig. 1, and the behaviour over different parameter dimensions and effective sparsity levels is shown in Fig. 2. The convergence differences with different noise cases are shown in Table IV. The emphasis of the comparison is on the effect of iteration numbers, parameter dimension, effective sparsity and stochastic noise on optimization behavior. Based on the results, the factors are related to the convergence of the three existing optimization methods present. stationary region. SGD, Adam, and AMSGrad are compared and contrasted for convergence, while ensuring the same initialization, dimension, sparsity and amount of noise.

VII. Conclusion

This study provided a convergence analysis that unifies SGD, Adam, and AMSGrad for high dimensional nonconvex machine-learning optimization. It accounted for various factors such as the dimension of the parameters, the sparsity of the effective gradient, stochastic gradient noise, and adaptive parameters on the behaviour of optimization.

The main convergence criterion employed was the expected squared gradient norm. The influence of the number of active coordinates on convergence in high-dimensional ambient dimension and effective gradient sparsity s was

analyzed. Coordinate-wise learning rate and its convergence with different stochastic conditions were also explored in Adam and AMSGrad's adaptive update mechanisms.

These three established optimization methods were tested through an experimental framework with varying dimensional, sparsity and noise conditions. The results empirically validate in an empirical way, the theoretical relationships between convergence, parameter dimension, effective sparsity and stochastic gradient noise.

The study offers an overall theoretical and experimental basis to analyze existing optimization methods for their use in high-dimensional machine-learning applications. The introduction of a new optimizer is not discussed; instead, the analysis is done to have an understanding of convergence and stability characteristics of SGD, Adam and AMSGrad. Further research could involve scaling the analysis up to more complex sparsity masks, more complicated stochastic-gradient requirements and larger machine-learning models.

References

1. Zhou, D., Chen, J., Cao, Y., Yang, Z., & Gu, Q. (2024). On the convergence of adaptive gradient methods for nonconvex optimization. *Transactions on Machine Learning Research*.

2. Ahn, K., & Cutkosky, A. (2024). Adam with model exponential moving average is effective for nonconvex optimization. *Advances in Neural Information Processing Systems*.

3. Li, H., Rakhlin, A., & Jadbabaie, A. (2024). Convergence of Adam under relaxed assumptions. *Advances in Neural Information Processing Systems*.

4. Li, H., & Lin, Z. (2024). On the $O\left(\sqrt{d}/T^{1/4}\right)$convergence rate of RMSProp and its momentum extension measured by $\ell_1$norm. *arXiv preprint arXiv:2402.00389*.

5. Liu, Y., Pan, R., & Zhang, T. (2024). AdaGrad under anisotropic smoothness. *arXiv preprint arXiv:2406.15244*.

6. Maes, L., Zhang, T. H., Jolicoeur-Martineau, A., Mitliagkas, I., Scieur, D., Lacoste-Julien, S., & Guille-Escuret, C. (2024). Understanding Adam requires better rotation-dependent assumptions. *arXiv preprint arXiv:2410.19964*.

7. Maladkar, D., Jiang, R., & Mokhtari, A. (2024). Convergence analysis of adaptive gradient methods under refined smoothness and noise assumptions. *arXiv preprint arXiv:2406.04592*.

8. Taniguchi, S., Harada, K., Minegishi, G., Oshima, Y., Jeong, S. C., Nagahara, G., Iiyama, T., Suzuki, M., Iwasawa, Y., & Matsuo, Y. (2024). ADOPT: Modified Adam can converge with any β at the optimal rate. *International Conference on Learning Representations*.

9. Jiang, W., Yang, S., Wang, Y., & Zhang, L. (2024). Adaptive variance reduction for stochastic optimization under weaker assumptions. *Advances in Neural Information Processing Systems*.

10. Sato, N., Izumi, K., & Iiduka, H. (2024). Convergence analysis of adaptive optimization methods for nonconvex learning. *Journal of Machine Learning Research, 25*, 1–37.

11. Gong, X., Hao, J., & Liu, M. (2025). On the convergence of Adam-type algorithms for bilevel optimization under unbounded smoothness. *arXiv preprint*.

12. Crawshaw, M., & Liu, M. (2025). Complexity lower bounds of adaptive gradient algorithms for non-convex stochastic optimization under relaxed smoothness. *International Conference on Learning Representations*.

13. Zhang, Y., Zhao, D., & Li, H. (2025). AdamRAG: Adaptive algorithm with ravine method for training deep neural networks. *Neural Processing Letters, 57*, 53.

14. Jiang, W., Yang, S., Wang, Y., & Zhang, L. (2024). Adaptive variance reduction methods for stochastic optimization with weaker assumptions. *NeurIPS 2024*.

15. Ahn, K., & Cutkosky, A. (2024). Coordinate-wise adaptivity and convergence of Adam for nonconvex optimization. *NeurIPS 2024*.

16. Kunstner, F., Yadav, R., Milligan, A., Schmidt, M., & Bietti, A. (2025). Heavy-tailed class imbalance and why Adam outperforms gradient descent on language models. *International Conference on Learning Representations*.

17. Levy, K., Kavis, A., & Cevher, V. (2021). Storm+: Fully adaptive SGD with recursive momentum for nonconvex optimization. *Advances in Neural Information Processing Systems, 34*.

18. Zhou, D., Chen, J., Cao, Y., Yang, Z., & Gu, Q. (2024). Adaptive gradient methods for nonconvex optimization with improved dimension dependence. *Transactions on Machine Learning Research*.

19. Gong, X., Hao, J., & Liu, M. (2025). On the convergence of Adam-type optimization methods under unbounded smoothness. *arXiv preprint*.

20. Jiang, W., Yang, S., Wang, Y., & Zhang, L. (2024). Adaptive stochastic optimization and variance reduction under weaker assumptions. *Advances in Neural Information Processing Systems*.