*Research Article*

# An Improvement of Stochastic Gradient Descent Approach for Mean-Variance Portfolio Optimization Problem

## Stephanie S. W. Su [ID] and Sie Long Kek [ID]

*Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia, Pagoh Campus 84600, Johor, Malaysia*

Correspondence should be addressed to Sie Long Kek; slkek@uthm.edu.my

In this paper, the current variant technique of the stochastic gradient descent (SGD) approach, namely, the adaptive moment estimation (Adam) approach, is improved by adding the standard error in the updating rule. The aim is to fasten the convergence rate of the Adam algorithm. This improvement is termed as Adam with standard error (AdamSE) algorithm. On the other hand, the mean-variance portfolio optimization model is formulated from the historical data of the rate of return of the S&P 500 stock, 10-year Treasury bond, and money market. The application of SGD, Adam, adaptive moment estimation with maximum (AdaMax), Nesterov-accelerated adaptive moment estimation (Nadam), AMSGrad, and AdamSE algorithms to solve the mean-variance portfolio optimization problem is further investigated. During the calculation procedure, the iterative solution converges to the optimal portfolio solution. It is noticed that the AdamSE algorithm has the smallest iteration number. The results show that the rate of convergence of the Adam algorithm is significantly enhanced by using the AdamSE algorithm. In conclusion, the efficiency of the improved Adam algorithm using the standard error has been expressed. Furthermore, the applicability of SGD, Adam, AdaMax, Nadam, AMSGrad, and AdamSE algorithms in solving the mean-variance portfolio optimization problem is validated.

## 1. Introduction

Recently, the application of the stochastic gradient descent (SGD) approach to machine learning and deep learning is actively explored. Moreover, due to the ability of the SGD approach in handling the stochastic optimization problems [1] and for solving optimization problems under the uncertainty environment [2, 3], the SGD approach and its variants have been developed rapidly. By virtue of this, the mean-variance portfolio optimization problem [4], which deals with risk and return, has attracted the attention of the investment community. The optimal decision on the portfolio selection is necessarily needed, where the scientific approach is employed in maximizing the return with the minimum risk [5]. However, this optimal decision is difficult to be made in advance.

In this paper, the disadvantage of the SGD approach, which is the slow convergence [6, 7], is noticed. To improve this weakness, the standard error from the sampling theory is added to the updating rule of the adaptive moment estimation (Adam) algorithm [8], which is the current variant of the SGD approach. On this basis, the convergence rate of the Adam algorithm is improved significantly. This improved version is then known as Adam with standard error (AdamSE) algorithm. On the other hand, the application of SGD methods, including Adam, adaptive moment estimation with maximum (AdaMax), Nesterov-accelerated adaptive moment estimation (Nadam), AMSGrad, and AdamSE approaches, for solving the mean-variance portfolio optimization problem is further studied. For this purpose, the historical data of the rate of return for the S&P 500 stock, 10-year Treasury bond, and money market are employed. Then, the mean-variance portfolio optimization model is formulated. During the calculation procedure, the iterative solution converges to the optimal portfolio solution, and the performance of these algorithms is presented.

The rest of the paper is organized as follows. In Section 2, the mean-variance portfolio optimization problem is described, where the expected return and the covariance matrix are expressed. In Section 3, the enhancement of the convergence rate of the Adam algorithm by using the standard error from the sampling theory is further discussed. The calculation procedure for the SGD, Adam, AdaMax, Nadam, AMSGrad, and AdamSE algorithms is summarized. In Section 4, a mean-variance portfolio optimization model is formulated using the historical data of S&P 500 stock, 10-year Treasury bond, and money market. Then, the model is solved by using the algorithms discussed, and the results are presented. Finally, some concluding remarks are made.

## 2. Problem Description

Consider a general mean-variance portfolio optimization problem for $n$ risky assets given by

$$\text{minimize } f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{\Sigma}\mathbf{w},$$

$$\text{subject to} \quad \mathbf{w}^T\mathbf{\mu} = R,$$

$$\mathbf{w}^T\mathbf{1} = 1, \tag{1}$$

$$\mathbf{w} \geq \mathbf{0},$$

where $\mathbf{w} = (w_1, \ldots, w_n)^T$ is the vector of portfolio weights of the assets and $\mathbf{\Sigma}$ represents the covariance for the assets. Here, $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$ gives the variance of the portfolio, $\mathbf{\mu} = (\mu_1, \ldots, \mu_n)^T$ is the vector of portfolio return mean, and $\mathbf{1} = (1, \ldots, 1)^T$ is the vector with 1s elements. Note that the targeted expected return $R$ depends on the risk tolerance of investors.

Furthermore, by using the geometric mean, the portfolio return mean is computed from

$$\mu_i = \left(\prod_{t=1}^{T}\left(1 + r_{i,t}\right)\right)^{1/T} - 1, \tag{2}$$

where $r_{i,t}$ is the rate of return of asset $i$ at time $t = 1, \ldots, T$ and $\mu_i$ is the mean of the rate of return for the asset $i$, whereas the covariance matrix of assets is defined as

$$\mathbf{\Sigma} = \frac{1}{T}\sum_{i=1}^{T}\left(r_{i,t} - \mu_i\right)\left(r_{j,t} - \mu_j\right), \tag{3}$$

for $i, j = 1, \ldots, n$. The assumption of the mean-variance portfolio optimization defined in (1) is to minimize the risk of an investment that is represented by the variance, at the same time, to satisfy the targeted return of the portfolio [9]. Since the risk is always related to randomness and uncertainty [10], the stochastic optimization approach will be used in solving the optimization problem defined in (1).

## 3. Stochastic Optimization Method

Now, let us define the Lagrange function as follows:

$$L(\mathbf{w},\mathbf{\lambda}) = \frac{1}{2}\mathbf{w}^T\mathbf{\Sigma}\mathbf{w} + \lambda_1\left(R - \mathbf{w}^T\mathbf{\mu}\right) + \lambda_2\left(1 - \mathbf{w}^T\mathbf{1}\right) + \mathbf{\lambda}_3^T\mathbf{w}, \tag{4}$$

where $\mathbf{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$ is the vector of the Lagrange multiplier. Then, the following first-order necessary conditions are derived:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{\Sigma}\mathbf{w} - \lambda_1\mathbf{\mu} - \lambda_2\mathbf{1} + \mathbf{\lambda}_3 = 0, \tag{5}$$

$$\frac{\partial L}{\partial \lambda_1} = R - \mathbf{w}^T\mathbf{\mu} = 0, \tag{6}$$

$$\frac{\partial L}{\partial \lambda_2} = 1 - \mathbf{w}^T\mathbf{1} = 0, \tag{7}$$

$$\mathbf{\lambda}_3^T\mathbf{w} = 0, \mathbf{\lambda}_3 \geq \mathbf{0}. \tag{8}$$

*3.1. Analytical Optimal Solution.* From (5), the optimal weighted value of the portfolio is calculated from

$$\mathbf{w}^* = \lambda_1\mathbf{\Sigma}^{-1}\mathbf{\mu} + \lambda_2\mathbf{\Sigma}^{-1}\mathbf{1}. \tag{9}$$

Refer to (6), the targeted expected return is provided by

$$R = \mathbf{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{w}^*. \tag{10}$$

Then, substitute (9) into (10) to have the targeted expected return in terms of the Lagrange multipliers, that is,

$$R = \lambda_1\mathbf{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{\mu} + \lambda_2\mathbf{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{1}. \tag{11}$$

Rewrite (7) to be

$$1 = \mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{w}^* \tag{12}$$

and substitute (9) into (12):

$$1 = \lambda_1\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{\mu} + \lambda_2\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}. \tag{13}$$

From (11) and (13), after doing some algebraic manipulations, the Lagrange multipliers are computed from

$$\lambda_1 = \frac{\left(\mathbf{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{1}\right) - R\left(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}\right)}{\left(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{\mu}\right)\left(\mathbf{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{1}\right) - \left(\mathbf{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{\mu}\right)\left(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}\right)}, \tag{14}$$

$$\lambda_2 = \frac{R\left(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{\mu}\right) - \left(\mathbf{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{\mu}\right)}{\left(\mathbf{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{1}\right)\left(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{\mu}\right) - \left(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}\right)\left(\mathbf{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{\mu}\right)}, \tag{15}$$

and $\lambda_3 = \mathbf{0}$. Therefore, from the discussion above, the analytical solution of the mean-variance portfolio optimization problem defined in (1), which is given by (9), (14), and (15), is resulted. However, this analytical solution is assumed to be not available due to the uncertainty and randomness of the variables.

### 3.2. Stochastic Gradient Descent Algorithm.

Referring to the mean-variance portfolio optimization problem defined in (1), let us introduce an augmented objective function $f_a(\mathbf{w})$ as

$$f_a(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{w} + \lambda_1\left(R - \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}\right) + \lambda_2\left(1 - \mathbf{w}^{\mathrm{T}}\mathbf{1}\right) + \boldsymbol{\lambda}_3^{\mathrm{T}}\mathbf{w}, \tag{16}$$

with $\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}}$. Since the existence of the uncertainty, the augmented objective function $f_a(\mathbf{w})$ defined in (16) can be rewritten as the expected objective function, given by

$$E\left[f_a(\mathbf{w})\right] = \frac{1}{n}\sum_{i=1}^{n} f_{a,i}(\mathbf{w}), \tag{17}$$

where $f_{a,i}(\mathbf{w})$, $i = 1, \ldots, n$, is the element of the augmented objective function $f_a(\mathbf{w})$ that is uniformly sampled at random and $E[\cdot]$ is the expectation operator. By virtue of this, the sampled gradient $\nabla f_{a,i}(\mathbf{w})$ is denoted as an unbiased estimator to be

$$E\left[\nabla f_a(\mathbf{w})\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_{a,i}(\mathbf{w}). \tag{18}$$

Notice that the first-order necessary condition for (16) is equivalent to the first-order necessary condition (5). That is,

$$\frac{\partial f_a(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial L(\mathbf{w}, \boldsymbol{\lambda})}{\partial \mathbf{w}}. \tag{19}$$

For convenience, define the stochastic gradient

$$g(\mathbf{w}) = \frac{\partial L}{\partial \mathbf{w}}, \tag{20}$$

which can be calculated from (5). The updating rule of the SGD approach is given by

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha \cdot g_j\left(\mathbf{w}^{(k)}\right), \tag{21}$$

with the step size $\alpha$, which is also known as the learning rate, $k$ is the number of iterations, and the random index $j$ is the gradient referred.

Hence, the calculation procedure of the SGD algorithm is summarized as in Algorithm 1.

### 3.3. Adaptive Moment Estimation Algorithm.

In the Adam approach [11], the exponential decaying averages of past gradients $m_k$ and past squared gradients $v_k$ are considered as follows:

$$m_k = \beta_1 m_{k-1} + (1 - \beta_1)g_k, \tag{22}$$

$$v_k = \beta_2 v_{k-1} + (1 - \beta_2)g_k^2, \tag{23}$$

where $g_k$ is the gradient, $\beta_1$ and $\beta_2$ are the decay rates, which are close to 1. Notice that $m_k$ and $v_k$ are estimates of the first moment (the mean) and the second moment (the uncentered variance) of the gradients, respectively. These biases are counteracted by using the bias-corrected first- and second-moment estimates, given by

$$\widehat{m}_k = \frac{m_k}{1 - \beta_1^k}, \tag{24}$$

$$\widehat{v}_k = \frac{v_k}{1 - \beta_2^k}. \tag{25}$$

Thus, Adam updating rule has been presented as follows:

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha \cdot \frac{\widehat{m}_k}{\sqrt{\widehat{v}_k} + \delta}, \tag{26}$$

where $\delta$ is the smoothing term used to avoid division by zero.

The calculation procedure of the Adam algorithm is summarized as in Algorithm 2.

### 3.4. Adaptive Moment Estimation with Maximum.

AdaMax, which is the adaptive moment estimation with maximum [11], is a variant of the Adam optimizer that uses the infinity ($\ell_\infty$) norm, while the Adam optimizer itself uses the $\ell_2$-norm for optimization. When generalizing the Adam algorithm to the $\ell_\infty$-norm, and hence in AdaMax, the gradient update is the maximum between the past gradients and current gradient, which is shown as

$$\begin{aligned} v_k &= \beta_2^\infty v_{k-1} + \left(1 - \beta_2^\infty\right)\left|g_k\right|^\infty \\ &= \max\left(\beta_2 \cdot v_{k-1}, \left|g_k\right|\right). \end{aligned} \tag{27}$$

Then, the updating rule of AdaMax is

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \frac{\alpha}{v_k} \cdot \widehat{m}_k. \tag{28}$$

The calculation procedure of the AdaMax algorithm is summarized as in Algorithm 3.

### 3.5. Nesterov-Accelerated Adaptive Moment Estimation.

Nadam, which is the Nesterov-accelerated adaptive moment estimation, combines Adam and NAG, which is the Nesterov acceleration gradient [12]. The Nadam algorithm is employed for noisy gradients or gradients with high curvatures. The NAG algorithm allows performing a more accurate step in the gradient direction by updating the parameters with the momentum step before computing the gradient. The learning process is accelerated by summing up the exponential decay of the moving averages for the previous and current gradients. It is resulting in a little faster training time than the Adam algorithm. Its updating rule is shown by

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \frac{\alpha}{\sqrt{\widehat{v}_k} + \delta}\left(\beta_1\widehat{m}_{k-1} + \frac{1 - \beta_1}{1 - \beta_1^k} \cdot g_k\right). \tag{29}$$

The calculation procedure of the Nadam algorithm is summarized as in Algorithm 4.

### 3.6. AMSGrad.

In setting, where the Adam algorithm converges to a suboptimal solution, it has been observed that some minibatches provide large and informative gradients, but as these minibatches only occur rarely, exponential

Data: given the initial value $\mathbf{w}^{(0)} = \mathbf{w}_0$, the number of samples $n$, the step size $\alpha$, and the tolerance $\varepsilon$. Set $k = 0$.
Step 1: evaluate the augmented objective function from (16).
Step 2: compute the stochastic gradient from (20).
Step 3: set the random index $j$.
Step 4: update the vector $\mathbf{w}^{(k)}$ from (21). If $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| < \varepsilon$, then stop the iteration. Otherwise, set $k = k + 1$, and repeat from Step 1.
**Remark**:
The tolerance is $\varepsilon = 10^{-6}$, and the learning rate is $\alpha = 0.001$.

ALGORITHM 1: SGD algorithm.

Data: given the initial value $\mathbf{w}^{(0)} = \mathbf{w}_0$, the number of samples $n$, the step size $\alpha$, and the tolerance $\varepsilon$. Set $k = 0$.
Step 1: evaluate the augmented objective function from (16).
Step 2: compute the stochastic gradient from (20).
Step 3: set the random index $j$.
Step 4: compute the decaying averages of past and past squared gradients from (22) and (23).
Step 5: calculate the bias-corrected first- and second-moment estimates from (24) and (25).
Step 6: update the vector $\mathbf{w}^{(k)}$ from (26). If $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| < \varepsilon$, then stop the iteration. Otherwise, set $k = k + 1$, and repeat from Step 1.
**Remark**:
The default values for the decay rates are $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the smoothing term is $\delta = 10^{-8}$, while the tolerance is $\varepsilon = 10^{-6}$, and the learning rate is $\alpha = 0.001$.

ALGORITHM 2: Adam algorithm.

Data: given the initial value $\mathbf{w}^{(0)} = \mathbf{w}_0$, the number of samples $n$, the step size $\alpha$, and the tolerance $\varepsilon$. Set $k = 0$.
Step 1: evaluate the augmented objective function from (16).
Step 2: compute the stochastic gradient from (20).
Step 3: set the random index $j$.
Step 4: compute the decaying averages of past and past squared gradients from (22) and (27).
Step 5: calculate the bias-corrected first-moment estimate from (24).
Step 6: update the vector $\mathbf{w}^{(k)}$ from (28). If $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| < \varepsilon$, then stop the iteration. Otherwise, set $k = k + 1$, and repeat from Step 1.
**Remark**:
The default values for the decay rates are $\beta_1 = 0.9$ and $\beta_2 = 0.999$, the tolerance is $\varepsilon = 10^{-6}$, and the learning rate is $\alpha = 0.001$.

ALGORITHM 3: AdaMax algorithm.

averaging diminishes their influence, which leads to poor convergence. To fix this behaviour, a new algorithm, which is known as the AMSGrad algorithm [13], that uses the maximum of past squared gradients rather than the exponential average to update the parameters is created:

$$\widehat{v}_k = \max\left(\widehat{v}_{k-1}, v_k\right). \tag{30}$$

Therefore, the updating rule of AMSGrad is

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \frac{\alpha}{\sqrt{\widehat{v}_k} + \delta} \cdot m_k. \tag{31}$$

The calculation procedure of the AMSGrad algorithm is summarized as Algorithm 5.

### 3.7. Improved Adaptive Moment Estimation Algorithm.

Consider the standard error (SE) from the sampling theory [14],

$$\text{SE} = \frac{\sigma}{\sqrt{n}}, \tag{32}$$

where $\sigma$ is the population standard deviation and $n$ is the sample size for the sampling. Thus, for improving the updating rule of the Adam algorithm, assume that the standard error of the bias-corrected first-moment estimate $\widehat{m}_k$ is defined by

$$\widehat{s}_k = \frac{\sqrt{\widehat{v}_k} + \delta}{\sqrt{k}}, \tag{33}$$

where $\sqrt{\widehat{v}_k}$ represents the sample standard deviation of the gradient $g_k$ and $k$ is the number of iterations. From (26), the updating rule of the Adam algorithm is modified to be

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha \cdot \frac{\widehat{m}_k}{\widehat{s}_k}. \tag{34}$$

Data: given the initial value $\mathbf{w}^{(0)} = \mathbf{w}_0$, the number of samples $n$, the step size $\alpha$, and the tolerance $\varepsilon$. Set $k = 0$.
Step 1: evaluate the augmented objective function from (16).
Step 2: compute the stochastic gradient from (20).
Step 3: set the random index $j$.
Step 4: compute the decaying averages of past and past squared gradients from (22) and (23).
Step 5: calculate the bias-corrected first- and second-moment estimates from (24) and (25).
Step 6: update the vector $\mathbf{w}^{(k)}$ from (29). If $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| < \varepsilon$, then stop the iteration. Otherwise, set $k = k + 1$, and repeat from Step 1.
**Remark**:
The default values for the decay rates are $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the smoothing term is $\delta = 10^{-8}$, while the tolerance is $\varepsilon = 10^{-6}$, and the learning rate is $\alpha = 0.001$.

ALGORITHM 4: Nadam algorithm.

Data: given the initial value $\mathbf{w}^{(0)} = \mathbf{w}_0$, the number of samples $n$, the step size $\alpha$, and the tolerance $\varepsilon$. Set $k = 0$.
Step 1: evaluate the augmented objective function from (16).
Step 2: compute the stochastic gradient from (20).
Step 3: set the random index $j$.
Step 4: compute the decaying averages of past and past squared gradients from (22) and (23).
Step 5: calculate the bias-corrected moment estimate based on (30).
Step 6: update the vector $\mathbf{w}^{(k)}$ from (31). If $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| < \varepsilon$, then stop the iteration. Otherwise, set $k = k + 1$, and repeat from Step 1.
**Remark**:
The default values for the decay rates are $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the smoothing term is $\delta = 10^{-8}$, while the tolerance is $\varepsilon = 10^{-6}$, and the learning rate is $\alpha = 0.001$.

ALGORITHM 5: AMSGrad algorithm.

For the modification made, this improved Adam algorithm is also known as Adam with standard error (AdamSE) algorithm [11, 14].

The calculation procedure for the AdamSE algorithm is summarized as Algorithm 6.

## 4. Illustrative Example

Consider a portfolio optimization problem [15], where the portfolio selection is based on three securities, namely, S&P 500 stock, 10-year Treasury bond, and money market (MM). The corresponding historical data of the annual rate of return for these securities, which are dated from 1961 to 2003, are shown in Table 1.

By using (2) and (3), the mean of the return and the related covariance of this portfolio selection are calculated, and their values are, respectively, given by

$$\mu = \begin{pmatrix} 0.1073 \\ 0.0737 \\ 0.0627 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 0.02778 & 0.00387 & 0.00021 \\ 0.00387 & 0.01112 & -0.00020 \\ 0.00021 & -0.00020 & 0.00115 \end{pmatrix}.$$

(35)

Consequently, the mean-variance portfolio optimization model is formulated as follows:

$$\text{minimize} f(w_1, w_2, w_3) = \frac{1}{2}\left(0.02778w_1^2 + 2(0.00387)w_1w_2 + 2(0.00021)w_1w_3 + 0.01112w_2^2 \right.$$
$$\left. + 2(-0.00020)w_2w_3 + 0.00115w_3^2\right),$$

$$\text{subject to} \quad 0.1073w_1 + 0.0737w_2 + 0.0627w_3 = R,$$

$$w_1 + w_2 + w_3 = 1,$$

$$w_1, w_2, w_3 \geq 0.$$

(36)

with the initial weight $w_1 = 0.3$, $w_2 = 0.3$, and $w_3 = 0.4$ and the targeted expected return. $R = 0.065$.

As a result, the optimal portfolio in percentage is shown in Table 2, where the final iterative solutions obtained from

Data: given the initial value $\mathbf{w}^{(0)} = \mathbf{w}_0$, the number of samples $n$, the step size $\alpha$, and the tolerance $\varepsilon$. Set $k = 0$.
Step 1: evaluate the augmented objective function from (16).
Step 2: compute the stochastic gradient from (20).
Step 3: set the random index $j$.
Step 4: compute the decaying averages of past and past squared gradients from (22) and (23).
Step 5: calculate the bias-corrected first- and second-moment estimates from (24) and (25).
Step 6: calculate the standard error of the bias-corrected first-moment estimate from (33).
Step 7: update the vector $\mathbf{w}^{(k)}$ from (34). If $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| < \varepsilon$, then stop the iteration. Otherwise, set $k = k + 1$, and repeat from Step 1.
**Remark**:
The default values for the decay rates are $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the smoothing term is $\delta = 10^{-8}$, while the tolerance is $\varepsilon = 10^{-6}$, and the learning rate is $\alpha = 0.001$ as the same as in the Adam algorithm.

ALGORITHM 6: AdamSE algorithm.

TABLE 1: Annual rate of return for the portfolio.

| Year | Stock | Bond | MM |
| --- | --- | --- | --- |
| 1961 | 26.81 | 2.20 | 2.33 |
| 1962 | −8.78 | 5.72 | 2.93 |
| 1963 | 22.69 | 1.79 | 3.38 |
| 1964 | 16.36 | 3.71 | 3.85 |
| 1965 | 12.36 | 0.93 | 4.32 |
| 1966 | −10.10 | 5.12 | 5.40 |
| 1967 | 23.94 | −2.86 | 4.50 |
| 1968 | 11.00 | 2.25 | 6.03 |
| 1969 | −8.47 | −5.63 | 8.96 |
| 1970 | 3.94 | 18.92 | 4.90 |
| 1971 | 14.30 | 11.24 | 4.14 |
| 1972 | 18.99 | 2.39 | 5.33 |
| 1973 | −14.69 | 3.30 | 9.95 |
| 1974 | −26.47 | 4.00 | 8.53 |
| 1975 | 37.23 | 5.52 | 5.20 |
| 1976 | 23.93 | 15.56 | 4.65 |
| 1977 | −7.16 | 0.38 | 6.56 |
| 1978 | 6.57 | −1.26 | 10.03 |
| 1979 | 18.61 | 1.26 | 13.78 |
| 1980 | 32.50 | −2.48 | 18.90 |
| 1981 | −4.92 | 4.04 | 12.37 |
| 1982 | 21.55 | 44.28 | 8.95 |
| 1983 | 22.56 | 1.29 | 9.47 |
| 1984 | 6.27 | 15.29 | 8.38 |
| 1985 | 31.73 | 32.27 | 8.27 |
| 1986 | 18.67 | 22.39 | 6.91 |
| 1987 | 5.25 | −3.03 | 6.77 |
| 1988 | 16.61 | 6.84 | 8.76 |
| 1989 | 31.69 | 18.54 | 8.45 |
| 1990 | −3.10 | 7.74 | 7.31 |
| 1991 | 30.46 | 19.36 | 4.43 |
| 1992 | 7.62 | 7.34 | 2.92 |
| 1993 | 10.08 | 13.06 | 2.96 |
| 1994 | 1.32 | −7.32 | 5.45 |
| 1995 | 37.58 | 25.94 | 5.60 |
| 1996 | 22.96 | 0.13 | 5.29 |
| 1997 | 33.36 | 12.02 | 5.50 |
| 1998 | 28.58 | 14.45 | 4.68 |
| 1999 | 21.04 | −7.51 | 5.30 |
| 2000 | −9.10 | 17.22 | 6.40 |
| 2001 | −11.89 | 5.51 | 1.82 |
| 2002 | −22.10 | 15.15 | 1.24 |
| 2003 | 28.68 | 0.54 | 0.98 |

TABLE 2: Optimal portfolio percentage.

| Solution | Stock | Bond | MM |
| --- | --- | --- | --- |
| Benchmark | 2.637097 | 10.226016 | 87.136887 |
| SGD | 2.630398 | 10.244025 | 87.125477 |
| Adam | 2.630400 | 10.244023 | 87.125478 |
| AdaMax | 2.630646 | 10.243939 | 87.125515 |
| Nadam | 2.630399 | 10.244023 | 87.125479 |
| AMSGrad | 2.630400 | 10.244022 | 87.125479 |
| AdamSE | 2.630361 | 10.244024 | 87.125707 |

TABLE 3: Performance of algorithms.

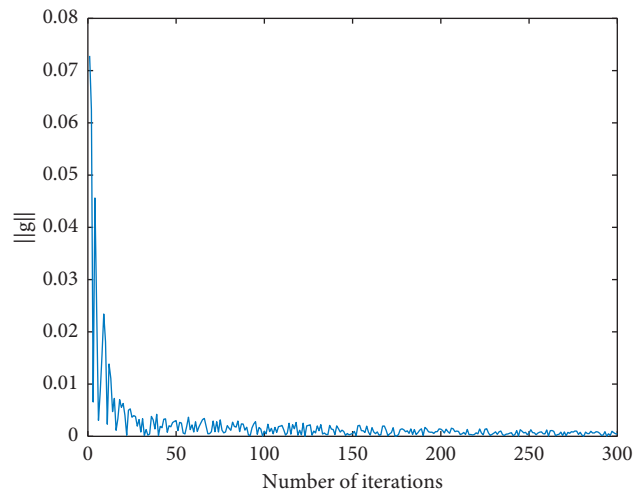| Algorithm | SGD | Adam | AdaMax | Nadam | AMSGrad | AdamSE |
|---|---|---|---|---|---|---|
| Number of iterations | 2265 | 2242 | 7208 | 2243 | 2273 | 303 |



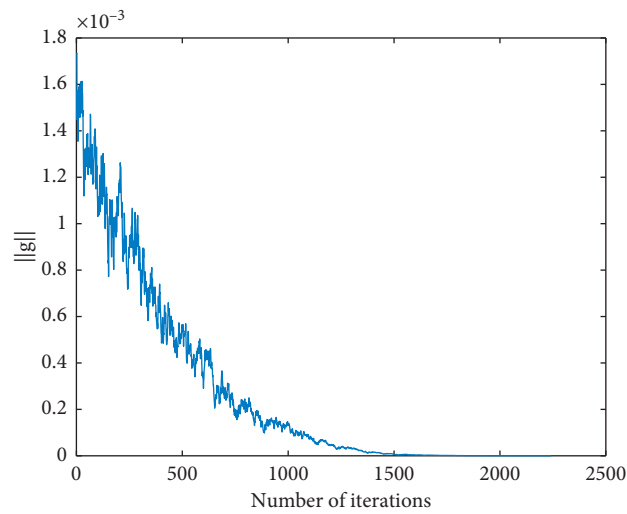FIGURE 1: Convergence behavior of SGD algorithm for the first 300 of 2265 iteration numbers.
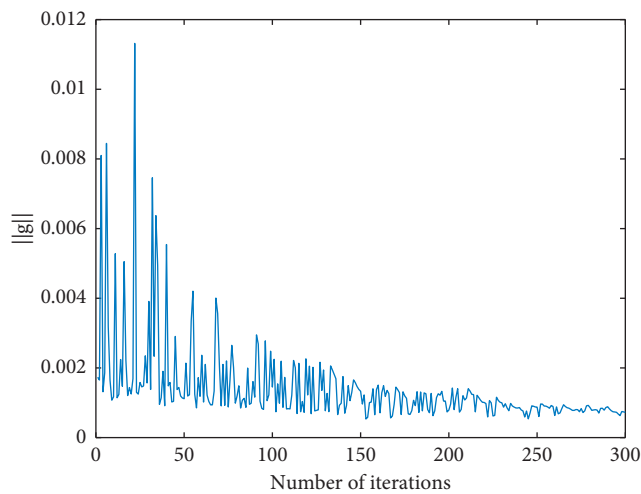


FIGURE 2: Convergence behavior of Adam algorithm.



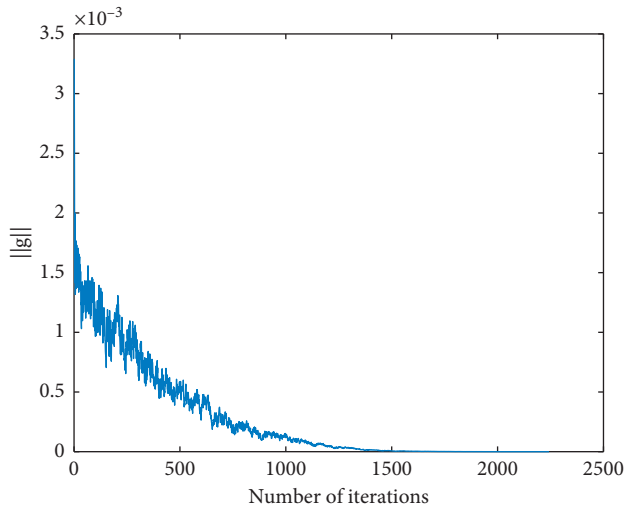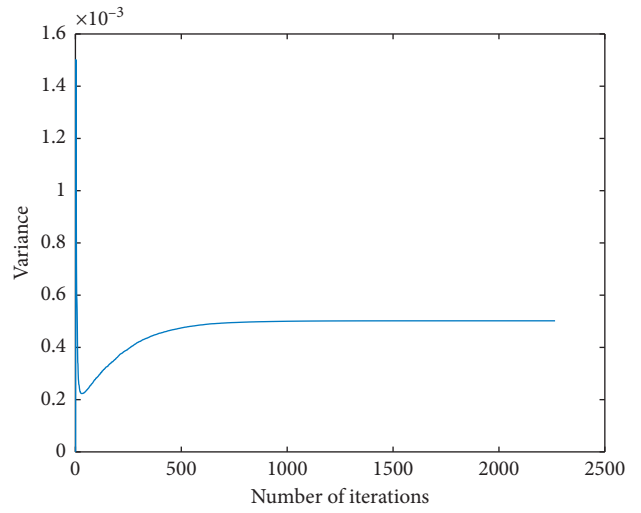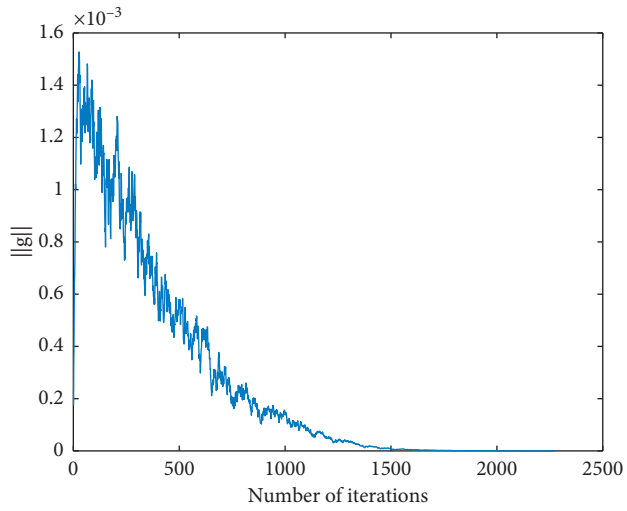FIGURE 3: Convergence behavior of AdaMax algorithm for the first 300 of 7208 iteration numbers.

FIGURE 4: Convergence behavior of Nadam algorithm.



FIGURE 7: Changes in the variance of SGD algorithm.



FIGURE 5: Convergence behavior of AMSGrad algorithm.
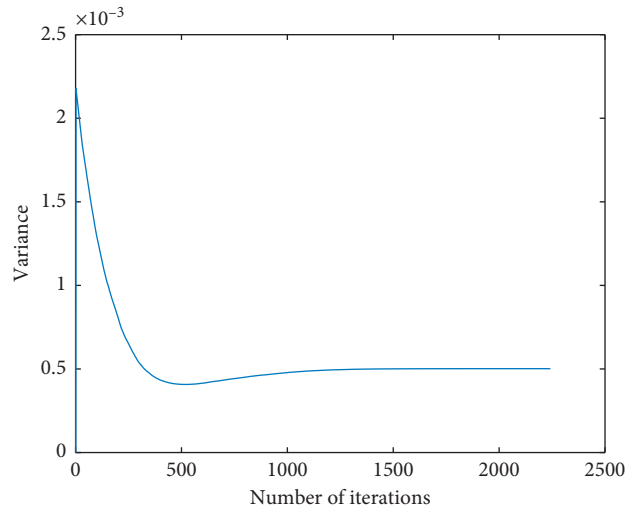


FIGURE 8: Changes in the variance of Adam algorithm.
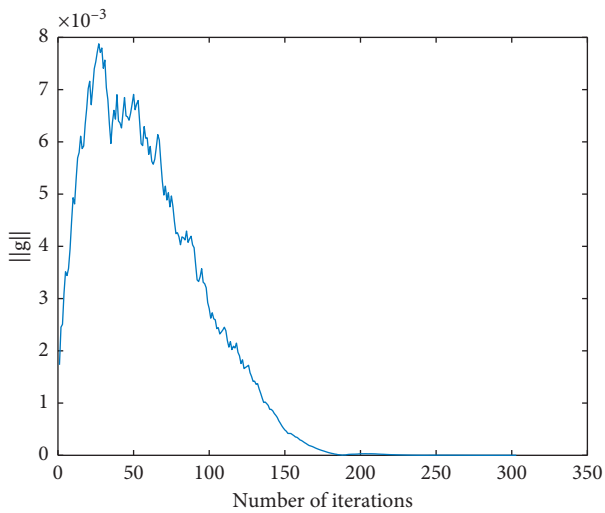


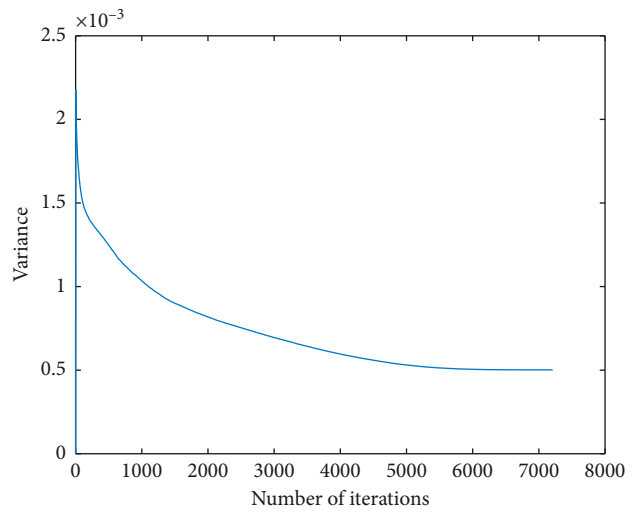FIGURE 6: Convergence behavior of AdamSE algorithm.



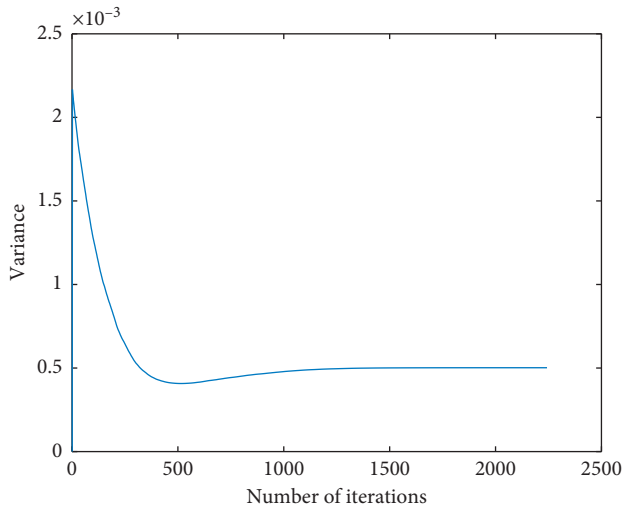FIGURE 9: Changes in the variance of AdaMax algorithm.

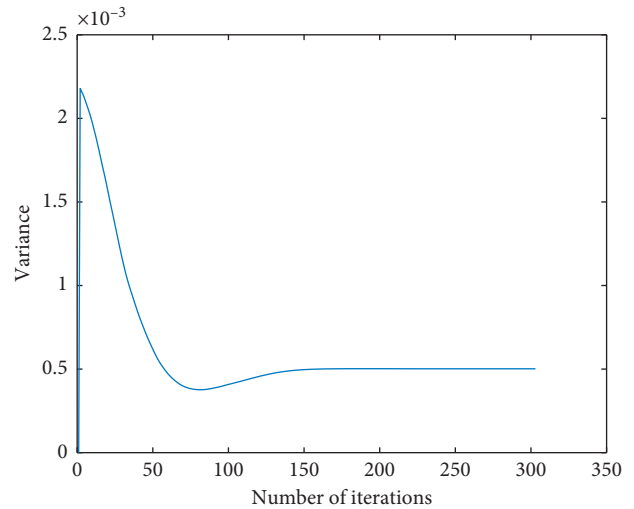FIGURE 10: Changes in the variance of Nadam algorithm.



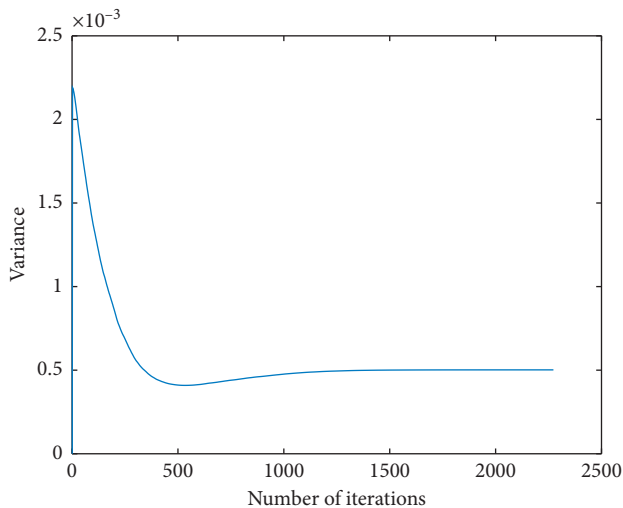FIGURE 12: Changes in the variance of AdamSE algorithm.



FIGURE 11: Changes in the variance of AMSGrad algorithm.

the SGD, Adam, AdaMax, Nadam, AMSGrad, and AdamSE algorithms are compared. The optimal solution from [15] is enclosed as the benchmark solution. It could be noticed that these algorithms are able to provide the optimal weight for the portfolio selection, which is given by (2.63, 10.24, 87.13). The Lagrange multipliers are $\lambda_1 = 7.2 \times 10^{-3}$ and $\lambda_2 = 5.3268 \times 10^{-4}$.

The performance of the respective methods in solving the mean-variance portfolio optimization problem is shown in Table 3, where the numbers of iterations are presented.

Apparently, the AdamSE algorithm has the smallest number of iterations, which is an 86 percent reduction from the iteration numbers of the Adam algorithm, while the Adam algorithm leads the iterative solution to the convergence with 1 percent faster than the SGD algorithm. At the same time, some variants of Adam algorithm, namely, AdaMax, Nadam, and AMSGrad, show more iteration numbers than the Adam algorithm. Also, the convergence behaviour for each algorithm, which is

represented by the norm of the stochastic gradient, is shown in Figures 1–6, respectively. For better visualization, the iterative results for the first 300 iteration numbers of the SGD and AdaMax algorithms are presented as shown in Figures 1 and 3, respectively. Therefore, the modification of the Adam algorithm by equipping the standard error significantly enhances the rate of the convergence of the Adam algorithm and the efficiency of the AdamSE algorithm is definitely proven.

In addition to this, the objective function has a minimum risk of $5.0182 \times 10^{-4}$, where the changes in the variance presented by each algorithm during the iteration procedure are shown in Figures 7–12, respectively. From these figures, it is noticed that the variance is dramatically increasing before meeting a peak point for the different algorithms, and then the variance is gradually decreasing and heading to the minimum variance of $5.0182 \times 10^{-4}$. This behaviour indicates that the divergence issue has been counted when reaching a peak point, and after satisfying the constraints, the optimal solution is successfully determined to give the optimal weight of the portfolio selection.

## 5. Concluding remarks

The enhancement of the convergence rate of the Adam algorithm, which is addressed by using the standard error, was discussed in this paper. This improved version of the algorithm is known as the AdamSE algorithm. In addition, the application of the algorithms of SGD, Adam, AdaMax, Nadam, AMSGrad, and AdamSE in solving the mean-variance portfolio optimization problem was also studied. The result obtained showed that the AdamSE algorithm is an efficient approach, especially for solving the mean-variance portfolio optimization problem. In conclusion, the practicality of the SGD algorithm and its current variants, which are the Adam, AdaMax, Nadam, AMSGrad, and AdamSE algorithms, is particularly validated for the mean-variance portfolio optimization problem.

## Data Availability

The data used are shown in Table 1.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## References

[1] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points: online stochastic gradient for tensor decomposition," *Journal of Machine Learning Research*, vol. 40, pp. 1–46, 2015.

[2] L. Rigotti and C. Shannon, "Uncertainty and risk in financial markets," *Econometrica*, vol. 73, no. 1, pp. 203–243, 2005.

[3] J. F. Moxnes and K. Hausken, "Introducing randomness into first-order and second-order deterministic differential equations," *Advances in Mathematical Physics*, vol. 2010, no. 1, 42 pages, Article ID 509326, 2010.

[4] A. Dangi, "Financial portfolio optimization: computationally guided agents to investigate, analyse and invest!?" 2013, https://arxiv.org/abs/1301.4194.

[5] K. T. Lwin, R. Qu, and B. L. MacCarthy, "Mean-VaR portfolio optimization: a nonparametric approach," *European Journal of Operational Research*, vol. 260, no. 2, pp. 751–766, 2017.

[6] Q. Tu, Y. Rong, and J. Chen, "Parameter identification of ARX models based on modified momentum gradient descent algorithm," *Complexity*, vol. 2020, p. 11, 2020.

[7] D. Driggs, M. J. Ehrhardt, and C. B. Schönlieb, "Accelerating variance-reduced stochastic gradient methods," *Mathematical Programming*, 2020.

[8] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2015, https://arxiv.org/abs/1412.6980.

[9] Y. C. Duan, "Multi-objective approach to portfolio optimization," *Rose-Hulman Undergraduate Mathematics Journal*, vol. 8, no. 1, pp. 1–18, 2007.

[10] T. S. Coleman, *Quantitative Risk Management: A Practical Guide to Financial Risk*, John Wiley & Sons, Hoboken, NJ, USA, 2012.

[11] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, pp. 1–15, San Diego, CA, USA, May 2015.

[12] T. Dozat, "Incorporating Nesterov momentum into Adam," in *Proceedings of the International Conference on Learning Representations Work*, vol. 1, pp. 1–4, Boston, MA, USA, April 2016.

[13] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *Proceedings of the international Conference on learning representations*, pp. 1–23, Boston, MA, USA, April 2018.

[14] G. Keller, *Statistics for Management and Economics Abbreviated*, South-Western Cengage Learning, Boston, MA, USA, 9th edition, 2012.

[15] G. Cornuejols and R. Tütüncü, *Optimization Methods in Finance*, Cambridge University Press, Pittsburgh, PA, USA, 2006.