Vietnam Journal of Mathematics
https://doi.org/10.1007/s10013-024-00688-w

ORIGINAL ARTICLE

Standard Error Adaptive Moment Estimation for Mean-Value-at-Risk Portfolio Optimization Problems by Sampling

Stephanie See Weng Su[1] · Sie Long Kek[1] · Kok Lay Teo[2]

Received: 5 May 2023 / Accepted: 30 January 2024

Abstract

In this paper, an improvement of the adaptive moment estimation (Adam) method equipped with standard error (SE), namely the AdamSE algorithm, is proposed. Our aims are to improve the convergence rate of the Adam algorithm and to explore the utility of the AdamSE algorithm for solving mean-value-at-risk (mean-VaR) portfolio optimization problems. For this, 10 stocks from the top 30 equity holdings list released by the Employees Provident Fund (EPF) have a weak correlation among them. The weekly stock prices of these stocks are selected for the period from 2015 to 2019, and then the mean, covariance and required rate of return are calculated to build a mean-VaR portfolio optimization model. In this way, the Adam and AdamSE algorithms are used to solve the model, and their results are compared. During the calculation, the stochastic gradients of the model are simulated through sampling, and nine samples are taken into consideration. With this sampling, the standard error of each sample is computed and the optimal weight for each sample is determined using the AdamSE algorithm. After convergence is achieved, the results show that different sample sizes could provide a satisfactory outcome for the portfolio concerned and from these nine samples, the lowest and highest iteration numbers were obtained to guarantee a robust optimal solution to the model constructed. Hence, we concluded that the AdamSE algorithm through sampling reveals its computational capability for handling the mean-VaR portfolio optimization problem.

We dedicate this research article to members and postgraduate students in our research group. This is to encourage them to continue mathematical research works in optimization. We appreciate Prof. Dr. Hoang Xuan Phu's invitation to contribute the article to this special issue dedicated to Prof. Dr. Tamás Terlaky on his birthday.

✉ Sie Long Kek
slkek@uthm.edu.my

Stephanie See Weng Su
hw150074@siswa.uthm.edu.my

Kok Lay Teo
koklayt@sunway.edu.my; K.L.Teo@curtin.edu.au

[1] Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia, Pagoh Campus, Pagoh 84600, Johor, Malaysia

[2] School of Mathematical Sciences, Sunway University, 47500 Kuala Lumpur, Malaysia

Published online: 01 April 2024

Keywords Stochastic gradient · Adaptive moment estimation · Standard error · Sampling · Mean-VaR portfolio optimization

Mathematics Subject Classification (2010) 62D05 · 62P05 · 90C15 · 90C20 · 91G10

1 Introduction

Adaptive moment estimation (Adam) method proposed in [10] is one of the latest stochastic gradient descent methods. This method is another adaptive learning rate method among other adaptive learning rate methods in the literature, see for examples, momentum [16], Nesterov accelerated gradient (NAG) [15], Adagrad [4], Adadelta [24], AdaMax [10], Nadam [3], AMSGrad [17], AdamW [12], QHAdam [14], and AggMo [13]. In Adam method, the decay averages of past gradients and past squared gradients are computed and stored to give estimates of the first and second moments of the gradients, respectively. These estimates are initialized to zeros vectors and are biased towards to zero, especially during the initial time steps and when the decay rates are small. These biases are counteracted by computing bias-corrected first and second moment estimates to update the parameters during the calculation procedure.

The Adam method has been extensively studied across multiple fields, such as risk management, portfolio selection, and machine learning. Schiele [19] improved the accuracy of the asset return estimation and the expected associated portfolio performance through a dynamic portfolio optimization framework and the artificial neural network. Ghahtarani et al. [6] reviewed recent robust portfolio selection problems from operational research and financial perspectives. From their study, the classification of models and methods was presented. Veraguas et al. [23] considered stochastic optimal control problems for which a risk minimization problem for controlled diffusions was solved. They derived a dynamic programming principle to recover central results of risk-neutral, and the value of the risk minimization problem can be characterized as a viscosity of a Hamilton–Jacobi–Bellman–Isaacs equation. Chronopoulos et al. [2] studied a deep quantile estimator based on a neural network to forecast value-at-risk (VaR) and to find significant gains over linear quantile regression, where the Adam algorithm is used to train the neural network.

On the other hand, portfolio optimization problems have been well-studied in economics and finance. Baltas et al. [1] studied a robust-entropic optimal control problem for portfolio management. They provided a closed-form solution and a detailed study of the limiting behaviour by associated stochastic differential game. Thus, the effect of robustness on the optimal decisions of both players was clarified. Temocin et al. [21] considered the optimal portfolio problem with minimum guarantee protection in a defined contribution pension scheme. They compared various versions of the guarantee concept, and each guarantee framework was obtained through a classical stochastic control approach. Kara et al. [9] considered the robust conditional VaR under parallelepiped uncertainty in modelling the robust optimal portfolio allocation. From their finding, the stability of portfolio allocation was increased, and the portfolio risk was reduced. Savku and Weber [18] discussed optimal investment problems using stochastic differential game approaches. They derived regime-switching Hamilton–Jacobi–Bellman–Isaacs equations to obtain explicit optimal portfolio strategies with Feynman–Kac representations of value functions.

In our study, we improve the convergence rate of the Adam method by reducing the iteration number. For this, the standard error (SE) is added to the updating rule of the Adam algorithm, and hence, the name AdamSE algorithm is given. To begin, a mean-value-at-risk (mean-VaR) portfolio optimization problem for the Employees Provident Fund (EPF) is formulated. The mean, covariance and required rate of return are calculated from the weekly stock prices of 10 assets selected for the period from 2015 to 2019. The simulation results obtained by using the Adam and AdamSE algorithms are compared and discussed. In addition, we consider nine samples of the past gradients through sampling simulation, which differs from only one sample in [20]. Therefore, different iteration numbers are given to arrive at the optimal weights and three different confidence levels are used to provide the portfolio risk for the model under study.

This paper is organized as follows. In Section 2, a mean-VaR portfolio optimization problem for the EPF is described. The weekly stock prices of 10 assets from the top 30 equity holdings list released by the EPF are utilized to calculate the mean, covariance and required rate of return. These parameters are used to construct the portfolio model. In Section 3, the Lagrange function is defined and the first order necessary conditions are derived. Furthermore, the calculation procedures of Adam and AdamSE algorithms are presented. In Section 4, simulation results obtained using the Adam and AdamSE algorithms are provided. In addition, the results of the nine samples through sampling are discussed. Finally, concluding remarks are given.

2 Problem Description

Consider a mean-VaR portfolio optimization problem [7, 25], which is to minimize the objective function,

$$f(w) = z_\alpha \sqrt{w^\top \Sigma w}\sqrt{\Delta t} \tag{1}$$

subject to the following constraints,

$$w^\top \mu = R, \tag{2}$$

$$w^\top I = 1, \tag{3}$$

$$0 \leq w \leq 1, \tag{4}$$

where $w = (w_1, w_2, \ldots, w_n)^\top \in \Re^n$ is an n-vector of the portfolio weights, $\Sigma \in \Re^{n\times n}$ is an $n \times n$ covariance matrix of the portfolio, and $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^\top \in \Re^n$ is an n-vector of the expected return rate of the portfolio, whereas $I = (1, 1, \ldots, 1)^\top \Re^n$ is an n-vector with 1s elements, and R is the minimum threshold at which investors can tolerate the expected rate of return on their portfolio.

Here, the portfolio's VaR is given by the objective function (1), the confidence level α reflects the degree of risk aversion, z_α is the z-score for the confidence level α and Δt is the holding period. Since the portfolio consists of a set of assets with uncertain stock prices, and the portfolio weights are random variables, for which the initial weights are average weight. This mean-VaR problem is defined as a stochastic optimization problem.

Now, a mean-VaR portfolio optimization problem is stated as follows. Consider the case where 10 stocks are selected [5] from the top 30 equity holdings list released by the EPF. These stock prices have a weak correlation. The weekly stock prices of these stocks are selected for the period from 2015 to 2019 and retrieved from the website investing.com. Using these past historical stock prices data, the mean, covariance and required rate of return of the portfolio are calculated, and they are given below.

(a) The means of return rates

$$\mu = \begin{pmatrix} -0.001935268 \\ -0.000349588 \\ 0.001131086 \\ -0.00147822 \\ 0.000463904 \\ 0.000831973 \\ -0.00354601 \\ -0.000959335 \\ -0.000252542 \\ 0.00302638 \end{pmatrix}.$$

(b) The covariance of the portfolio

$$\Sigma = 10^{-3} \times \begin{bmatrix} 1.192 & 0.151 & 0.297 & 0.339 & 0.106 & 0.329 & 0.198 & 0.388 & 0.213 & 0.213 \\ 0.151 & 1.094 & 0.072 & 0.205 & 0.108 & 0.143 & 0.217 & 0.375 & 0.138 & 0.136 \\ 0.297 & 0.072 & 2.805 & 0.210 & 0.091 & 0.240 & 0.394 & 0.261 & 0.116 & 0.276 \\ 0.339 & 0.205 & 0.210 & 1.594 & 0.197 & 0.307 & 0.248 & 1.004 & 0.327 & 0.215 \\ 0.106 & 0.108 & 0.091 & 0.197 & 0.339 & 0.190 & 0.139 & 0.193 & 0.131 & 0.013 \\ 0.329 & 0.143 & 0.240 & 0.307 & 0.190 & 1.299 & 0.268 & 0.407 & 0.260 & 0.134 \\ 0.198 & 0.217 & 0.394 & 0.248 & 0.139 & 0.268 & 1.804 & 0.646 & 0.144 & 0.300 \\ 0.388 & 0.375 & 0.261 & 1.004 & 0.193 & 0.407 & 0.646 & 2.829 & 0.285 & 0.188 \\ 0.213 & 0.138 & 0.116 & 0.327 & 0.131 & 0.260 & 0.144 & 0.285 & 0.537 & 0.216 \\ 0.213 & 0.136 & 0.276 & 0.215 & 0.013 & 0.134 & 0.300 & 0.188 & 0.216 & 0.851 \end{bmatrix}.$$

(c) The required return rate

$$R = 0.0005.$$

Here, the holding period Δt = 260 days, the confidence level α = 0.05 and the z-score z_α = 1.645 are used in the model's objective function (1).

Therefore, the mean-VaR portfolio optimization problem for the EPF investment is constructed by substituting the values of mean, covariance and required return rate into (1) and (2). Define the portfolio's weight $w = (w_1, w_2, \ldots, w_{10})^\top \in \Re^{10}$, which is regarded as a random variable vector, and the aim is to determine the optimal weight w for these 10 assets of the portfolio such that the VaR of the portfolio is minimized.

3 Adaptive Moment Estimation with Standard Error

Define the Lagrange function,

$$L(w, \lambda) = z_\alpha \sqrt{w^\top \Sigma w}\sqrt{\Delta t} + \lambda_1(R - w^\top \mu) + \lambda_2(1 - w^\top I) + \lambda_3^\top w, \tag{5}$$

where $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ is the multiplier vector with $\lambda_1, \lambda_2 \in \Re$ and $\lambda_3 \in \Re^n$ to be determined later. From (5), the first order necessary conditions for the model are derived as follows:

$$\frac{\partial L(w, \lambda)}{\partial w} = \frac{z_\alpha \sqrt{\Delta t}(\Sigma w)}{\sqrt{w^\top \Sigma w}} - \lambda_1 \mu - \lambda_2 I + \lambda_3 = 0, \tag{6}$$

$$\frac{\partial L(w,\lambda)}{\partial \lambda_1} = w^\top \mu - R = 0,$$
$$\frac{\partial L(w,\lambda)}{\partial \lambda_2} = w^\top I - 1 = 0,$$
$$\lambda_3^\top w = 0, \quad \lambda_3 \geq 0.$$

Here, (6) is the gradient of the mean-VaR model, which is employed in Adam and AdamSE algorithms to find the optimal weights for the mean-VaR portfolio optimization problem.

3.1 Analytical Solution for Deterministic Case

We now consider the case where the portfolio weight w is deterministic, requiring to be determined. Multiplying $w^\top$ to (6) and doing some algebraic manipulations, we obtain the standard deviation of the portfolio as follows,

$$\sqrt{w^\top \Sigma w} = \frac{\lambda_1 R + \lambda_2}{z_\alpha \sqrt{\Delta t}}. \tag{7}$$

Substituting (7) into (6), we obtain the weight as given below:

$$w = \Sigma^{-1} \frac{(\lambda_1 R + \lambda_2)(\lambda_1 \mu + \lambda_2 I - \lambda_3)}{z_\alpha^2 \Delta t}. \tag{8}$$

From (7) and (8), it follows that λ_1 and λ_2 are given by

$$\lambda_1 = \frac{AR - B}{\sqrt{(AC - B^2)(AR^2 - 2BR + C)}} 2 z_\alpha \sqrt{\Delta t}, \tag{9}$$
$$\lambda_2 = \frac{C - BR}{\sqrt{(AC - B^2)(AR^2 - 2BR + C)}} 2 z_\alpha \sqrt{\Delta t}, \tag{10}$$

with $A = I^\top \Sigma^{-1} I$, $B = I^\top \Sigma^{-1} \mu$, $C = \mu^\top \Sigma^{-1} \mu$ and $\lambda_3 = 0$.

According to the above discussion, the analytical solution of the mean-VaR portfolio optimization problem [25] defined by (1)–(4) is determined by (8), (9) and (10). However, we assume that the analytical solution does not exist since the weight variables are random variables, which depend on the availability of the expected return rate and the covariance of the portfolio. In addition, the stock prices in the portfolio are uncertain and random.

3.2 Adam Algorithm

Consider the exponential moving averages of past gradients m_k and past squared gradients v_k are, respectively, given as follows:

$$m_k = \beta_1 m_{k-1} + (1 - \beta_1) g_k, \tag{11}$$
$$v_k = \beta_2 v_{k-1} + (1 - \beta_2) g_k^2, \tag{12}$$

where the term g_k is the gradient at the time step k, parameter β_1 is the exponential decay rate for the first moment (mean) estimates of the gradient. In contrast, the parameter β_2 is the exponential decay rate for the second moment (uncentered variance) estimates of the gradient. Since the average of the past gradient m_k is the first moment, it resembles the momentum that records the past normalized gradients. While the squared gradient v_k is the second moment that gives different learning rates for different parameters.

The moment estimates are biased towards zero, especially during the initial time steps and low decay rates. These biases can be counteracted by using the bias-corrected first and second-moment estimates given by

$$\hat{m}_k = \frac{m_k}{1-\beta_1^k}, \tag{13}$$

$$\hat{v}_k = \frac{v_k}{1-\beta_2^k}. \tag{14}$$

When the moments m_k and v_k are expanded and expressed by the gradient g_k, it is found that after dividing by the correction factor $1-\beta^k$, the sum of the coefficients of all gradients g_i approximates to 1, so it is called the normalized correction. Both moments m_k and v_k are initialized to 0, the gradients have not accumulated in the first few iterations, and the values of moments m_k and v_k are close to 0. In particular, the parameter β_2 is often set closer to 1 than β_1, which leads to the initial update step size being too large. By normalization correction, the moments m_k and v_k can be enlarged so that the size of moment estimates $\hat{m}_k$ and $\hat{v}_k$ with a small k value is at the same level as that of moment estimates $\hat{m}_k$ and $\hat{v}_k$ when the gradient with a large k value has been fully accumulated.

The Adam algorithm updates exponential moving averages of past gradient m_k and past squared gradient v_k by using hyper-parameters $\beta_1, \beta_2 \in [0, 1)$ to control the exponential decay rates of the moving in (13) and (14). The Adam algorithm has the following updating rule,

$$w^{(k+1)} = w^{(k)} - \alpha_r \times \frac{\hat{m}_k}{\sqrt{\hat{v}_k} + \delta}, \tag{15}$$

where $\alpha_r, \delta > 0$. In the Adam algorithm, the learning rate that increases or decreases its value is dependent on the gradient value of the loss function. The learning rate will be lower for the higher gradient values, and the learning rate will be larger for the lower gradient values. Hence, the learning decelerates at steeper and speeds up at shallower parts of the loss function curve.

The learning rate for the Adam algorithm is set at $\alpha_r(\sqrt{\hat{v}_k} + \delta)^{-1}$. Its value varies from one iteration to another iteration because the parameter α_r is divided by the square root of the mean square sum of $(1-\beta_2)^{-1}$ parametric gradients at each iteration. The gradient of each parameter is different, so the learning rate of each parameter is not the same even in the same iteration. Moreover, the direction of parameter update is not only the gradient g_k of the current iteration but also the average of the gradient of the current and the past iterations, that is $(1-\beta_1)^{-1}$.

The parameter δ is a small number that prevents any division by zero during the algorithm implementation. Assuming $\delta = 0$, the effective step taken in the parameter space at iteration k is $\Delta_k = \alpha_r \times \hat{m}_k/\sqrt{\hat{v}_k}$, where the smaller signal-to-noise ratio (SNR), represented by $\hat{m}_k/\sqrt{\hat{v}_k}$, indicates that there is a greater uncertainty about whether the direction $\hat{m}_k$ corresponds to the direction of the actual gradient. Meanwhile, the effective step size is closer to zero towards an optimum when the SNR is small. This SNR is often close to zero, resulting in smaller effective steps in the parameter space. When approaching the minimum value, the noise in all directions will be very large, resulting in an SNR close to 0. The updating step size quickly reduces to 0, which is called automatic annealing, as mentioned in [10]. When a saddle point is encountered, the noise generated by moving around the saddle point can quickly make the current point jump out of the saddle point.

The calculation procedure of the Adam algorithm is summarized as Algorithm 1.

Algorithm 1 (Adam algorithm).

Data: Given the initial value $w^{(0)} = w_0$, the number of samples n, the step-size α_r and the tolerance ε. Set $k = 0$.

Step 1: Set the random index j.

Step 2: Evaluate the augmented objective function $L_j(w^{(k)}, \lambda)$ from (5).

Step 3: Compute the stochastic gradient $\nabla L_j(w^{(k)}, \lambda)$ from (6).

Step 4: Compute the decaying averages of the past gradient m_k and past squared gradient v_k from (11) and (12), respectively.

Step 5: Calculate the bias-corrected first and second-moment estimates, which are $\hat{m}_k$ and $\hat{v}_k$, from (13) and (14), respectively.

Step 6: Update the vector $w^{(k)}$ from (15). If $\|w^{(k+1)} - w^{(k)}\| < \varepsilon$ then stop the iteration. Otherwise, set $k = k + 1$, and repeat from *Step* 1.

Remark 1 The default values for the decay rates [10] are $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the smoothing term is $\delta = 10^{-8}$, while the tolerance is $\varepsilon = 10^{-6}$, and the learning rate is $\alpha = 0.001$.

3.3 AdamSE Algorithm

From the perspective of sampling theory, the standard error is used to measure the discrepancy of the sample mean and the population mean [11]. In other words, the standard error measures how accurately a sample distribution is representative of a population by using the standard deviation. The standard error is defined by

$$SE = \frac{\sigma}{\sqrt{n}}, \tag{16}$$

where σ is the population standard deviation and n is the sample size of the sampling distribution concerned. This standard error will increase when the population standard deviation increases, while this standard error will decrease when the sample size is increased. According to the central limit theorem [8], as the sample size approaches the actual population size, the sample means will increasingly cluster around the true population mean.

From the observation, we notice that the Adam algorithm uses the sampled gradients. Therefore, we assume that multiple gradient samples can be generated with a fixed sample size. From this point of view, we hypothesize that the standard error can be reduced and it is more appropriate to use instead of the standard deviation. This is because the standard error varies with sample size, but the standard deviation does not. Thus, to improve the updating rule of the Adam algorithm, it is assumed that the sampling distribution of the average past gradient m_k follows a normal distribution of the biased-corrected first and second moments $\hat{m}_k$ and $\hat{v}_k$, the standard error of the bias-corrected first-moment estimate $\hat{m}_k$, similar to (16), is defined as

$$\hat{s}_k = \frac{\sqrt{\hat{v}_k} + \delta}{\sqrt{n}}, \tag{17}$$

where n is the number of samples of the average past gradient m_k and δ is a very small positive number that prevents division by zero during the implementation. As the result of (17), the updating rule (15) of the Adam algorithm is replaced by

$$w^{(k+1)} = w^{(k)} - \alpha_r \times \frac{\hat{m}_k}{\hat{s}_k}, \tag{18}$$

as the updating rule of the AdamSE algorithm.

Note that the standard error is always smaller than the standard deviation [22]. Thus, with smaller standard errors, the step size of the AdamSE algorithm will become more effective than the step size of the Adam algorithm. This effective step size of the AdamSE algorithm speeds up the optimal search step. Therefore, we express the result in the following theorem.

Theorem 1 *Suppose that the step size of the AdamSE algorithm is*

$$\Delta_k = \alpha_r \times \frac{\hat{m}_k}{\hat{s}_k}. \tag{19}$$

Then, the convergence rate of the AdamSE algorithm is better than the convergence rate of the Adam algorithm. That is,

$$\|w^{(k+1)} - w^{(k)}\|_{AdamSE} \leq \|w^{(k+1)} - w^{(k)}\|_{Adam}. \tag{20}$$

Proof From (17) and (19), consider the step size of the AdamSE algorithm,

$$\Delta_k = \alpha_r \times \frac{\hat{m}_k}{\hat{s}_k} = \alpha_r \times \frac{\hat{m}_k}{\frac{\sqrt{\hat{v}_k}+\delta}{\sqrt{n}}} = \alpha_r \times \frac{\hat{m}_k}{\sqrt{\hat{v}_k}+\delta} \cdot \frac{1}{\sqrt{n}} \leq \frac{\hat{m}_k}{\sqrt{\hat{v}_k}+\delta}.$$

Therefore, the convergence rate of the AdamSE algorithm follows from (20) for $n \geq 1$. This completes the proof. □

The calculation procedure for the AdamSE algorithm is summarised as Algorithm 2.

Algorithm 2 (AdamSE algorithm).

```
Data: Given the initial value $w^{(0)} = w_0$, the number of samples $n$, the step-size $\alpha_r$ and the tolerance $\varepsilon$. Set $k = 0$.
Step 1: Set the random index $j$.
Step 2: Evaluate the augmented objective function $L_j(w^{(k)}, \lambda)$ from (5).
Step 3: Compute the stochastic gradient $\nabla L_j(w^{(k)}, \lambda)$ from (6).
Step 4: Compute the decay averages of the past gradient $m_k$ and past squared gradient $v_k$ from (11) and (12), respectively.
Step 5: Calculate the bias-corrected first and second-moment estimates, which are $\hat{m}_k$ and $\hat{v}_k$, from (13) and (14), respectively.
Step 6: Calculate the standard error of the bias-corrected first-moment estimate $\hat{s}_k$ from (17).
Step 7: Update the vector $w^{(k)}$ from (18). If $\|w^{(k+1)} - w^{(k)}\| < \varepsilon$, then stop the iteration. Otherwise, set $k = k + 1$ and repeat from Step 1.
```

Remark 2 The default values for the decay rates [10] are $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the smoothing term is $\delta = 10^{-8}$, while the tolerance is $\varepsilon = 10^{-6}$, and the learning rate is $\alpha = 0.001$. These values are the same as in the Adam algorithm.

4 Illustrative Results

The optimal portfolio weights after implementing the Adam and AdamSE algorithms are shown in Table 1, where only one sample ($n = 1$) of the past gradients is employed in the AdamSE algorithm.

Moreover, from Table 2, the AdamSE algorithm takes 43 number of iterations to converge, which is 76.8% faster than the Adam algorithm with the number of iterations being 185. From

Table 1 Optimal portfolio weights

Stock	Adam	AdamSE
1	0.0538	0.0435
2	0.0853	0.0897
3	0.0263	0.0236
4	0.0017	0.0031
5	0.4964	0.5111
6	0.0121	0.0189
7	0.0155	0.0182
8	0.0017	0.0032
9	0.1478	0.1290
10	0.1594	0.1597

Table 2 Performance of algorithms

	Adam	AdamSE
Number of iterations	185	43

Table 3 Optimal portfolio weights for different sample sizes

	Sample number								
Stock	1	2	3	4	5	6	7	8	9
1	0.0435	0.0541	0.0538	0.0537	0.0536	0.0540	0.0538	0.0539	0.0533
2	0.0897	0.0919	0.0917	0.0929	0.0901	0.0925	0.0918	0.0913	0.0908
3	0.0236	0.0271	0.0265	0.0250	0.0256	0.0260	0.0261	0.0260	0.0257
4	0.0031	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0062
5	0.5111	0.4971	0.4967	0.4968	0.4950	0.4937	0.4927	0.4953	0.4893
6	0.0189	0.0114	0.0117	0.0129	0.0119	0.0106	0.0119	0.0110	0.0114
7	0.0182	0.0162	0.0158	0.0151	0.0158	0.0154	0.0156	0.0156	0.0154
8	0.0032	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0062
9	0.1290	0.1386	0.1395	0.1410	0.1452	0.1458	0.1459	0.1446	0.1447
10	0.1597	0.1602	0.1609	0.1592	0.1594	0.1586	0.1588	0.1589	0.1570

Table 4 Performance of AdamSE algorithm

Sample numbers	1	2	3	4	5	6	7	8	9
Number of iterations	43	115	130	132	135	128	125	123	126

Table 5 Portfolio risk

Confidence level (%)	z-Score	Portfolio risk (%)
90	1.282	29.78
95	1.645	38.21
99	2.326	54.03

this result, we can see that these two algorithms are able to give the same optimal weights for the mean-VaR model. This shows that the AdamSE algorithm performs as well as the Adam algorithm in providing an optimal solution to the mean-VaR portfolio optimization problem.

Table 3 shows the simulation results when we consider different sample sizes of past gradients for $n = 1, 2, 3, 4, 5, 6, 7, 8, 9$ using the AdamSE algorithm. Although more samples could be considered, we only have a limited number of samples in this simulation to avoid any unnecessary problems such as divergence.

In addition, the performance of the AdamSE algorithm (measured in terms of the number of iterations for these sample sizes) is shown in Table 4. From the theoretical results, the smaller the number of iterations, the faster the algorithm converges. However, when using the AdamSE algorithm, there is no linear relationship between the number of samples and the number of iterations because the number of iterations decreases after a sample size of 5. Thus, the optimal weights are robust solutions that are not affected by the number of iterations.

The portfolio risk of the mean-VaR model under different confidence levels is shown in Table 5. We only consider 90%, 95% and 99% confidence levels, where the portfolio risk is increased when the confidence level increases. This indicates that the portfolio investment of the EPF becomes riskier as the confidence level increases, which may increase the possibility of causing the maximum loss on the investment.

5 Concluding Remarks

This paper discussed the improvement of the Adam algorithm, adding the standard error in the updating rule of the Adam algorithm. The aim is to improve the convergence rate of the Adam algorithm, so the improved algorithm is named AdamSE algorithm. For illustration, the mean-VaR portfolio optimization problem for the EPF was formulated. This portfolio optimization problem was solved using Adam and AdamSE algorithms, giving rise to their respective optimal weights. These two optimal weights turn out to be the same. However, the AdamSE algorithm took fewer number of iterations to converge. In our study, past gradients of nine samples were simulated through sampling. Different iteration numbers showed robust optimal weights. From these results, we concluded that the AdamSE algorithm is an efficient algorithm for handling the mean-VaR portfolio optimization problem. For future research, the practicality of the AdamSE algorithm will be investigated for solving nonlinear stochastic optimization problems.

Acknowledgements This research was supported by the Universiti Tun Hussein Onn Malaysia (UTHM) through Tier 1 (vot Q121), and the Ministry of Higher Education (MoHE) Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2021/STG06/SYUC/03/1).

Author Contributions The first and third authors contributed equally to this work.

References

1. Baltas, I., Xepapadeas, A., Yannacopoulos, A.N.: Robust portfolio decisions for financial institutions. J. Dyn. Games **5**, 61–94 (2018)
2. Chronopoulos, I., Raftapostolos, A., Kapetanios, G.: Forecasting value-at-risk using deep neural network quantile regression. J. Financ. Econom. (2023). https://doi.org/10.1093/jjfinec/nbad014
3. Dozat, T.: Incorporating Nesterov momentum into Adam. In: Proceedings of the 4th International Conference on Learning Representations, pp. 1–4 (2016)
4. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
5. Employees Provident Fund (EPF): Top 30 equity holdings by percentage of issued shares as at 31 March 2020. https://www.kwsp.gov.my/-/list-of-top-30-equity-holdings-by-percentage-of-issued-shares (2020)
6. Ghahtarani, A., Saif, A., Ghasemi, A.: Robust portfolio selection problems: a comprehensive review. Oper. Res. Int. J. **22**, 3203–3264 (2022)
7. Guo, X., Chan, R.H., Wong, W.-K., Zhu, L.X.: Mean–variance, mean–VaR, and mean–CVaR models for portfolio selection with background risk. Risk Manag. **21**, 73–98 (2019)
8. Islam, M.R.: Sample size and its role in Central Limit Theorem (CLT). Int. J. Phys. Math. **1**, 37–47 (2018)
9. Kara, G., Özmen, A., Weber, G.-W.: Stability advances in robust portfolio optimization under parallelepiped uncertainty. Cent. Eur. J. Oper. Res. **27**, 241–261 (2019)
10. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. The 3rd International Conference on Learning Representations, pp. 1–15 (2015). arXiv:1412.6980 (2014)
11. Lee, D.K., In, J., Lee, S.: Standard deviation and standard error of the mean. Korean J. Anesth. **68**, 220–223 (2015)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings the 7th International Conference on Learning Representations, pp. 1–19 (2019)
13. Lucas, J., Sun, S., Zemel, R., Grosse, R.: Aggregated momentum: stability through passive damping. In: Proceedings of the 7th International Conference on Learning Representations, pp. 1–22 (2019)
14. Ma, J., Yarats, D.: Quasi-hyperbolic momentum and Adam for deep learning. In: Proceedings of the 7th International Conference on Learning Representations, pp. 1–38 (2019)
15. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $O(1/K^2)$. Sov. Math. Dokl. **269**, 543–547 (1983)
16. Qian, N.: On the momentum term in gradient descent learning algorithms. Neural Netw. **12**, 145–151 (1999)
17. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of Adam and beyond. In: Proceedings of the 6th International Conference on Learning Representations, pp. 1–23 (2018)
18. Savku, E., Weber, G.-W.: Stochastic differential games for optimal investment problems in a Markov regime-switching jump-diffusion market. Ann. Oper. Res. **312**, 1171–1196 (2022)
19. Schiele, P.: Modern approaches to dynamic portfolio optimization. Jr. Manag. Sci. **6**, 149–189 (2021)
20. Su, S.S.W., Kek, S.L.: An improvement of stochastic gradient descent approach for mean-variance portfolio optimization problem. J. Math. **2021**, 8892636 (2021)
21. Temocin, B.Z., Korn, R., Selcuk-Kestel, A.S.: Constant proportion portfolio insurance in defined contribution pension plan management. Ann. Oper. Res. **266**, 329–348 (2018)
22. Tölgyesi, C., Pénzes, Z.: Biostatistics. The University of Szeged, Szeged (2018). https://eta.bibl.u-szeged.hu/1920/1/Biostatistics.pdf
23. Veraguas, J.B., Reppen, A.M., Tangpi, L.: Stochastic control of optimized certainty equivalents. SIAM J. Financ. Math. **13**, 745–772 (2022)
24. Zeiler, M.D.: ADADELTA: An adaptive learning rate method. arXiv:1212.5701 (2012)
25. Zhou, S., Shi, B., Wen, Z.: Analysis of mean-VaR model for financial risk control. Syst. Eng. Procedia **4**, 40–45 (2012)