COMETa2026
8th INTERNATIONAL SCIENTIFIC CONFERENCE
12th - 14th November 2026
Jahorina, B&H, Republic of Srpska
University of East Sarajevo
Faculty of Mechanical Engineering
Conference on Mechanical Engineering Technologies and Applications

RECURSIVE IDENTIFICATION OF HAMMERSTEIN MODEL USING MULTI-INNOVATION HEAVY-BALL ALGORITHM WITH ADAPTIVE RESTART

Vojislav Filipović[1], Vladimir Đorđević[2]

Abstract: Optimisation is important in recent developments in data science and has applications in analysing and solving statistical and machine learning problems. High-dimensional parameter spaces characterise machine learning problems. As a result, first-order methods, particularly gradient-based algorithms, dominate machine learning optimisation. The main research direction is developing accelerated first-order methods. One such method, Polyak's heavy-ball method, is considered for identifying Hammerstein nonlinear systems with a memoryless nonlinear block and a linear dynamical block described by a DARMA (deterministic autoregressive moving average) model. To further improve the convergence rate of the heavy-ball algorithm, we extended the algorithm from the viewpoint of innovation modification. In the algorithm, we use the a posteriori prediction error. The fast gradient method relies on momentum to achieve an optimal convergence rate. However, momentum causes non-monotonic behaviour, leading to oscillations and overshooting in the function value (in our case, criterion identification). That is, the function values may increase for a couple of iterations before decreasing again. To avoid these problems, we introduce a restart mechanism in the algorithm. The main contributions of this paper are: a) design of a new multi-innovation form of the heavy-ball algorithm for identification, b) introduction of an adaptive restart strategy (function-based adaptive restart, gradient-based adaptive restart) to provide monotonicity of the heavy-ball algorithm.

Key words: Machine learning, Hammerstein model, Multi-innovations, Heavy-ball algorithm, Adaptive restart.

1 INTRODUCTION

Many practical modern problems involve large data volumes, inherent uncertainty, unavoidable nonconvexity, and the need to solve problems in real time. In addition, in a key area of artificial intelligence, machine learning, which is closely connected to control theory, problems are often characterised by high-dimensional

[1] Prof. Dr, Vojislav Filipović, University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia, v.filipovic@mts.rs
[2] Dr, Vladimir Đorđević, University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia, djordjevic.v@mfkv.kg.ac.rs (CA)

parameter spaces. Owing to this fact, first-order methods, particularly gradient-based algorithms, are dominant [1-2]. A main research direction in modern machine learning optimisation is developing accelerated first-order methods. Originating from Polyak's heavy-ball methods [3-4] and Nesterov's work [5-6], accelerated first-order methods have become an important topic in both optimisation and machine learning.

In this paper, we briefly describe the iterative gradient algorithm and Polyak's heavy-ball algorithm for minimising convex, smooth functions. Using these results later in the paper, we will develop a recursive version of the heavy-ball algorithm.

Real systems are generally nonlinear. Mathematical models for such kinds of systems are different. In this paper, we assume the system is described by the Hammerstein model [7], which consists of a linear time-invariant block in the form of a DARMA model and a static nonlinear block. The nonlinear part of the Hammerstein model is a nonlinear function of a known basis. The stochastic Hammerstein model, when the disturbance has a non-Gaussian distribution (presence of outliers), is considered in [8-9].

To further improve the convergence rate of the heavy-ball algorithm, we introduce history innovations [10]. The new form of Polyak's algorithm has a faster convergence rate and better parameter estimation accuracy than the standard Polyak's algorithm.

One characteristic of fast-gradient methods (heavy-ball method) is that the iterates are not monotone. That is, the function values may increase for a couple of iterations before decreasing again. This is undesirable, and the literature has proposed monotone versions of different fast-gradient methods [11-12]. In [13], an algorithm that exploits the non-monotonicity to achieve faster convergence is considered. It is observed that if the algorithm is restarted when non-monotonic behaviour is detected, the restarted algorithm often proceeds in a good direction in terms of function-value progress. Many references discuss the restart fast-gradient method [14-16].

In this paper, we adopt results for the heavy-ball algorithm [3-4] for application in the field of identification. With the introduction of the multi-innovation concept and an adaptive restart strategy, the new algorithm achieves a higher convergence rate. The main contributions of the paper are

a) design of a new multi-innovation form of heavy-ball algorithm for identification of nonlinear Hammerstein model,
b) introduction of an adaptive restart strategy to provide monotonicity of the heavy-ball algorithm.

2 VECTOR FORM OF HAMMERSTEIN MODEL FOR SISO SYSTEMS

Figure 1 presents the general form of the Hammerstein model for SISO systems.

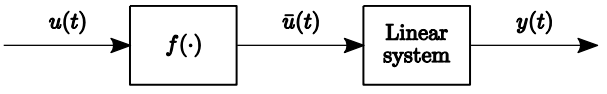

Figure 1. *The discrete-time Hammerstein model*

The nonlinear part of the Hammerstein model is a nonlinear function of known basis γ_1, γ_2, , γ_p [17]

$$\bar{u}(t)=\sum_{j=1}^{p}c_j\gamma_j\left(u(t)\right)=f\left(u(t)\right). \tag{1}$$

Let us suppose that the linear part of the Hammerstein model has a DARMA structure

$$A\left(q^{-1}\right)y(t)=B\left(q^{-1}\right)\bar{u}(t), \tag{2}$$

where

$$A\left(q^{-1}\right)=1+a_1q^{-1}+\ldots+a_nq^{-n},$$
$$B\left(q^{-1}\right)=b_1q^{-1}+\ldots+b_mq^{-m}.$$

Hammerstein model of the system is

$$A\left(q^{-1}\right)y(t)=B\left(q^{-1}\right)f\left(u(t)\right). \tag{3}$$

The transfer function of the linear system is

$$G\left(q^{-1}\right)=\frac{b_1q^{-1}+\ldots+b_mq^{-m}}{1+a_1q^{-1}+\ldots+a_nq^{-n}}. \tag{4}$$

The characterisation of the Hammerstein model in Figure 1, $f(u)$ and $G\left(q^{-1}\right)$ are not unique. In fact, any pair $(\alpha f(u), G\left(q^{-1}\right)/\alpha), \alpha\in(0,\infty)$ would produce identical input and output measurements. It is impossible to distinguish between $\left(f(u),\ G\left(q^{-1}\right)\right)$ and $\left(\alpha f(u),\ G\left(q^{-1}\right)/\alpha\right)$. We introduce the assumption that $c_1=1$, in relation (1), as in references [18-19].

From (2), it follows that

$$y(t)=-\sum_{i=1}^{n}a_iy(t-i)+\sum_{i=1}^{m}b_i\bar{u}(t-i)=-\sum_{i=1}^{n}a_iy(t-i)+\sum_{i=1}^{m}b_i\sum_{j=1}^{p}c_j\gamma_j\left(u(t-i)\right). \tag{5}$$

Let us define the parameter vector θ and information vector $\varphi(t)$ as

$$\theta=\begin{bmatrix}a\\c_1b\\\vdots\\c_pb\end{bmatrix}\in\mathbb{R}^{n_0},\quad \varphi(t)=\begin{bmatrix}\varphi_0(t)\\\varphi_1(t)\\\vdots\\\varphi_p(t)\end{bmatrix}\in\mathbb{R}^{n_0},\quad \varphi_0(t)=\begin{bmatrix}-y(t-1)\\-y(t-2)\\\vdots\\-y(t-n)\end{bmatrix}\in\mathbb{R}^{n}, \tag{6}$$

where $n_o=n+pm$, and

$$a=\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^n, \quad b=\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \in \mathbb{R}^m, \tag{7}$$

$$\varphi_j(t)=\begin{bmatrix} \gamma_j\left(u(t-1)\right) \\ \gamma_j\left(u(t-2)\right) \\ \vdots \\ \gamma_j\left(u(t-m)\right) \end{bmatrix} \in \mathbb{R}^m, \quad j=1,2,\ldots,p. \tag{8}$$

Finally, the vector form of the Hammerstein model is

$$y(t)=\varphi^T(t)\theta. \tag{9}$$

3 MULTI-INNOVATION GRADIENT ALGORITHM

In this part of the paper, we will consider the gradient descent algorithm. Convexity plays a fundamental role in this context [6]. A scalar function $f:\mathbb{R}^n \to \mathbb{R}$ is convex if

$$f\left(\lambda x+(1-\lambda)y\right) \le \lambda f(x)+(1-\lambda)f(y), \tag{10}$$

for all $(x,y)\in\mathbb{R}^n$ and $\lambda \in [0,1]$. We will consider the next optimisation problem

$$\min_x f(x), \tag{11}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ is a convex and smooth function. For the solution of problem (11), it is possible to use the following gradient descent algorithm

$$x_{k+1}=x_k-\rho\nabla f(x_w), \tag{12}$$

where $\rho>0$ is the step size. Let us suppose that the gradient of $f(x)$ satisfies the Lipschitz condition

$$\left\|\nabla f(x)-\nabla f(y)\right\| \le L\|x-y\|, \tag{13}$$

where $L>0$ is the Lipschitz constant. For $\rho\in(0,2/L)$ algorithm (12) is convergent.

For system (9), the a priori predicted output is

$$\hat{y}^o(t)=\varphi^T(t)\hat{\theta}(t-1), \tag{14}$$

where $\varphi(t)$ is given by relation (6) and

$$\hat{\theta}^T(t-1)=\left[\hat{a}_1(t-1),\ldots,\hat{a}_n(t-1),\hat{b}_n(t-1),\ldots,\hat{b}_m(t-1)\right]. \tag{15}$$

The a posteriori predicted output is

$$\hat{y}(t) = \varphi^T(t)\hat{\theta}(t). \tag{16}$$

The a priori prediction error is defined as

$$e^o(t) = y(t) - \hat{y}^o(t) = \varphi^T(t)\theta - \varphi^T(t)\hat{\theta}(t-1) = \varphi^T(t)\tilde{\theta}(t-1), \tag{17}$$

where $\tilde{\theta}(t-1) = \theta - \hat{\theta}(t-1)$. The a posteriori prediction error is

$$e(t) = y(t) - \hat{y}(t) = \varphi^T(t)\theta - \varphi^T(t)\hat{\theta}(t) = \varphi^T(t)\tilde{\theta}(t), \tag{18}$$

where $\tilde{\theta}(t) = \theta - \hat{\theta}(t)$.

Further, we will derive the multi-innovation gradient algorithm. We expand the scalar innovation $e(t)$ [10]

$$E(l,t) = \begin{bmatrix} e(t) \\ e(t-1) \\ \vdots \\ e(t-l+1) \end{bmatrix}, \tag{19}$$

where $E(l,t) \in \mathbb{R}^l$ and $l > 0$ denotes the innovation length and

$$e(t-i) = y(t-1) - \varphi^T(t-i)\hat{\theta}(t-i), \quad i = 0,1,2,\ldots,l. \tag{20}$$

In general, $\hat{\theta}(t-1)$ at time $t-1$ is closer to θ than $\hat{\theta}(t-i)$, $i = 2,3,\ldots,l-1$.

The innovation vector $E(l,t)$ has a form

$$E(l,t) = \begin{bmatrix} y(t) - \varphi^T(t)\hat{\theta}(t) \\ y(t-1) - \varphi^T(t-1)\hat{\theta}(t) \\ \vdots \\ y(t-l+1) - \varphi^T(t-l+1)\hat{\theta}(t) \end{bmatrix}. \tag{21}$$

Let us define the information matrix $\Phi(l,t) \in \mathbb{R}^{n_0 \times l}$ and stacked output vector $Y(l,t) \in \mathbb{R}^l$

$$\Phi(l,t) = \left[\varphi(t), \varphi(t-1), \ldots, \varphi(t-l+1)\right], \tag{22}$$

$$Y(l,t) = \left[y(t), y(t-1), \ldots, y(t-l+1)\right]. \tag{23}$$

The vector $E(l,t)$ now has a form

$$E(l,t) = Y(l,t) - \Phi^T(l,t)\hat{\theta}(t). \tag{24}$$

Now, the identification criterion is

$$J(\theta)=\frac{1}{2}\left\|E(l,t)\right\|^2=\frac{1}{2}\left\|Y(l,t)-\Phi^T(l,t)\theta\right\|^2, \tag{25}$$

The gradient algorithm, based on criterion (25), has a form

$$\hat{\theta}(t)=\hat{\theta}(t-1)-F\nabla_\theta J(\theta)=\hat{\theta}(t-1)+F\Phi(l,t)E(l,t), \tag{26}$$

where $F>0$ is the adaptation gain (learning rate) matrix. Two commonly used choices are

a) a general positive definite matrix $F>0$,

b) $F=\alpha I$, $\alpha\in(0,\infty)$ where I is the identity matrix.

A priori and a posteriori prediction output for the multi-innovation model are

$$\hat{Y}^o(l,t)=\Phi^T(l,t)\hat{\theta}(t-1), \tag{27}$$

$$\hat{Y}(l,t)=\Phi^T(l,t)\hat{\theta}(t). \tag{28}$$

Using relations (24), (26), (27) and (28), it follows that

$$\begin{aligned}E(l,t)&=Y(l,t)-\Phi^T(l,t)\hat{\theta}(t)=Y(l,t)-\Phi^T(l,t)\left[\hat{\theta}(t-1)+F\Phi^T(l,t)E(l,t)\right]=\\&=Y(l,t)-\Phi^T(l,t)\hat{\theta}(t-1)-\Phi^T(l,t)F\Phi(l,t)E(l,t)=\\&=E_o(l,t)-\Phi^T(l,t)F\Phi(l,t)E(l,t).\end{aligned} \tag{29}$$

From the last relation, we have

$$E(l,t)=\left[I+\Phi^T(l,t)F\Phi(l,t)\right]^{-1}E^o(l,t). \tag{30}$$

The last relation is important for implementing the recursive algorithm (26).

4 NEW FORM OF HEAVY-BALL RECURSIVE ALGORITHM

For the solution of problem (11), instead of algorithm (12), we use the heavy-ball (Polyak's) algorithm, which incorporates additional information (past estimate of the parameter). The heavy-ball is a two-step algorithm defined by the following state transitions

$$w_k=-\nabla f(x_k)+\beta w_{k-1}, \tag{31}$$

$$x_{k+1}=x_k+\alpha w_k. \tag{32}$$

for some initial points x_0 and w_0 and some positive sequences $\alpha\in(0,\infty)$ and $\beta\in[0,\infty)$. We will take $w_0=0$. We can rewrite the algorithm (31) and (32) in a different form.

From (32) we have

$$w_k = \frac{1}{\alpha}\left(x_k - x_{k-1}\right). \tag{33}$$

Then it follows that form (31)-(32)

$$x_{k+1} = x_k + \alpha\left(-\nabla f\left(x_k\right) + \frac{\beta}{\alpha}\left(x_k - x_{k-1}\right)\right). \tag{34}$$

Finally, from (34) follows

$$x_{k+1} = x_k - \alpha\nabla f\left(x_k\right) + \beta\left(x_k - x_{k-1}\right). \tag{35}$$

The last relation represents Polyak's algorithm and $x_k - x_{k-1}$ is referred to as momentum. The iteration (35) is equivalent to a discretisation of the second-order differential equation

$$\ddot{x} + a\dot{x} + b\nabla f\left(x\right) = 0, \tag{36}$$

which models the motion of a body in a potential field given by f with friction.

In the relation (35), the inertial term $\beta\left(x_k - x_{k-1}\right)$ incorporates information from the previous iteration and may significantly accelerate convergence. The parameters α and β can be selected as in [3-4]

$$0 \le \beta < 1, \quad 0 < \alpha < \frac{2\left(1+\beta\right)}{L}. \tag{37}$$

Using relations (25), (26), (30), and (35), one can get the new form of the heavy-ball algorithm based on the multi-innovation concept

$$\hat{\theta}\left(t\right) = \left(1+\beta\right)\hat{\theta}\left(t-1\right) - \beta\hat{\theta}\left(t-2\right) + F\Phi\left(l,t\right)E\left(l,t\right). \tag{38}$$

Remark 1. The connection between Kaczmarz and Polyak's heavy-ball algorithms is considered in [20].

Remark 2. A potential application of the heavy-ball method is nonlinear system identification, as described in reference [21].

5 RESTARTING OF HEAVY-BALL ALGORITHMS

First-order methods are well known to converge quickly at first, but convergence slows as iterations progress. This suggests that periodically restarting first-order methods could accelerate convergence. Figure 2 illustrates this concept. Let us suppose that

$$f^* = \min_x f\left(x\right). \tag{39}$$

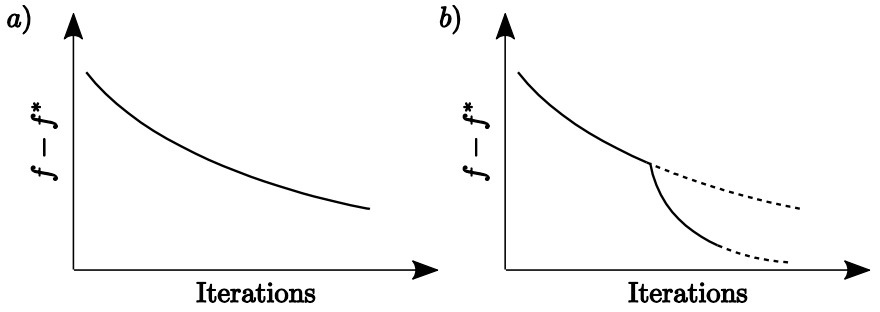

Figure 2. *a) Convergence plot without restart; b) Convergence plot with restart*

The first form of restart is a fixed restart. It means we restart the algorithm every k iterations, using the last point produced by the algorithm as the starting point and resetting the momentum to zero. The drawback of using fixed restart is the unknown k number of iterations.

The above analysis suggests that adaptive restart may be somewhat useful. For that, we want a scheme that makes some computationally cheap observations and decides whether to restart based on the observations. We will apply the function scheme [13] to our problem. The new form of the heavy-ball algorithm with adaptive restart, using relations (24), (30), and (38), is

$$\hat{\theta}(t)=\left(1+\beta(t-1)\right)\hat{\theta}(t-1)-\beta(t-1)\hat{\theta}(t-2)+F\Phi(l,t)\left[I+\Phi^{T}(l,t)F\Phi(l,t)\right]^{-1}E^{o}(l,t), \quad (40)$$

$$\beta(t-1)=\begin{cases}0, & E^{o}(l,t-1)>E^{o}(l,t-2)\\ \beta, & E^{o}(l,t-1)<E^{o}(l,t-2)\end{cases}. \quad (41)$$

Convergence analysis of algorithm (40) and (41) can be performed from the viewpoint of control theory. Namely, algorithm (40) and (41) can be presented as a Lur'e dynamic system [22] and convergence can be proven using LaSalle's invariance principle.

6 CONCLUSION

This paper presents preliminary results on a new form of the heavy-ball algorithm based on a multi-innovation strategy and adaptive restarts. This first-order algorithm is very important in machine learning and control theory. Further investigation is based on intensive simulations and the extension of results to multivariable and time-varying systems. Analysing the convergence of the proposed algorithm is also very important.

ACKNOWLEDGMENT

The Ministry of Science, Technological Development and Innovation of the Republic of Serbia supported this study through grant 451-03-34/2026-03/200108.

REFERENCES

[1] Liu, Z., Li, H., Fang, C. (2020*). Accelerated Optimization for Machine Learning*, Springer, Singapore.

[2] Franci, B., Grammatico, S. (2022). Convergence of sequences: A survey, *Annual Reviews in Control***, 53**, p.p. 161–186.

[3] Polyak, B. (1964). Some methods of speeding up the convergence of iteration methods, *USSR Computational Mathematics and Mathematical Physics***,** 4/5, p.p. 1–17.
[4] Polyak, B. (1987*). Introduction to Optimization*, Optimization Software, New York.
[5] Nesterov, Y. (1983). "A method for unconstrained convex minimization problem with the rate of convergence O ($1/k^2$)", *Dokl. Akad. Nauk. SSSR***, 269/3**, p.p. 543–547.
[6] Nesterov, Y. (2018*). Lectures of Convex Optimization*, Springer, Switzerland.
[7] Filipović, V. (2016*). Industrial Regulators (in Serbian)*, University of Kragujevac, Kraljevo.
[8] Filipović, V. (2017). Outlier robust stochastic approximation algorithm for identification of MIMO Hammerstein models, *Nonlinear Dynamics***, 90**, p.p. 1427–1441.
[9] Filipović, V. (2022). Outlier robust identification of dual-rate Hammerstein models in the presence of unmodeled dynamics, *Int. J. Robust Nonlinear Control*, **32**, p.p. 1162–1179.
[10] Ding, F., Chen, T. (2007). Performance analysis of multi-innovation gradient type identification methods, *Automatica***,** 43, p.p. 1–14.
[11] Back, A., Tebonlle, R. (2009). Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems, *IEEE Transactions on Image Processing***, 18/11**, p.p. 2419–2434.
[12] Nesterov, Y. (2005). Smooth minimization of non-smooth functions, *CIRP Annals - Mathematical Programming***,** 103/1, p.p. 127–152.
[13] O'Donoghue, B., Candes, E. (2015). Adaptive restart for accelerated gradient schemes, *Foundations of Computational Mathematics***,** 15/3, p.p. 715–732.
[14] Giselsson, P., Boyd, S. (2014). Monotonicity and restart in fast gradient methods, *53rd IEEE Conference on Decision and Control***,** p.p. 5058–5063.
[15] Kim, D., Fessler, J. (2018). Adaptive restart of the optimized gradient method for convex optimization, *Journal of Optimization Theory and Applications***,** 178/1, p.p. 240–263.
[16] d'Aspremont, A., Scieur, D., Taylor, A. (2021). Acceleration methods, *Foundations and trends in optimization***,** 5/1-2, p.p. 1–245.
[17] Cerone, V., & Regruto, D. (2003). Parameter bounds for discrete-time Hammerstein models with bounded output errors, *IEEE Transactions on Automatic Control***,** 48/10, p.p. 1855–1860.
[18] Bai, E. W. (1998). An optimal two-stage identification algorithm for Hammerstein–Wiener nonlinear systems, *Automatica***,** 34/3, p.p. 333–338.
[19] Gallman, P. (1976). A comparison of two Hammerstein model identification algorithms, *IEEE Transactions on Automatic Control***,** 21/1, p.p. 124–126.
[20] Wu, N. C., Zuo, Q., Wang, Y. (2024). On the Polyak momentum variants of the greedy deterministic single and multiple row-action methods, *Numerical Linear Algebra with Applications***,** 31/4, p.p. e2552.
[21] Mania, H., Jordan, M. I., Recht, B. (2022). Active learning for nonlinear system identification with guarantees, *Journal of Machine Learning Research***,** 23, p.p. 1–30.
[22] Heuts, Y. J., Padilla, G. P., Donkers, M. C. F. (2021). An adaptive restart heavy-ball projected primal-dual method for solving constrained linear quadratic optimal control problems, *2021 60th IEEE Conference on Decision and Control (CDC)***,** p.p. 6722–6727.