An Empirical Analysis of Convergence Dynamics for Quasi-Hyperbolic Momentum Methods

INTRODUCTION

Stochastic gradient descent's (SGD) utility as an algorithm in mathematical optimization and machine learning hardly needs an explanation. Yet despite its widespread use, momentum methods inspired by SGD have not been fully understood in both theory and practice. While mathematically rigorous, vanilla SGD suffers primarily from slow convergence, especially in the context of nonconvex function spaces such as deep learning [1]. To address slow convergence, numerous variants of SGD have been introduced as 'momentum-based methods' which implement additional parameters into the model to speed up convergence rates and stabilize optimization performance. One such approach, Quasi-Hyperbolic momentum (QHM), generalizes such momentum methods by introducing a parameter ν that interpolates between SGD and SHB along with a momentum parameter β that accelerates the optimization process and dampens oscillatory effects. Such a framework makes QHM very suitable for the analysis of momentum.

By combining theoretical insights with empirical experiments, we aim to offer clearer guidelines for choosing momentum parameters in practice. This paper builds on the insights from the article *Understanding the Role of Momentum in Stochastic Gradient Methods* [2] to provide an empirical validation of some key theoretical claims concerning the behavior of momentum parameters under QHM. Specifically, the paper exploits the QHM framework due to its generalizability in order to:

- Validate theoretical predictions regarding the behavior of QHM when the momentum parameter β approaches 1.
- Investigate the non-linear dependence of convergence rates on the parameters β and ν.
- Compare QHM's performance with other momentum-based algorithms such as Newton's-Accelerated Gradient (NAG) and Stochastic-Heavy Ball (SHB) in traditional machine learning tasks.

BACKGROUND

2.1 Momentum-Based Stochastic Gradient Methods

Briefly, SGD is a classical minimization problem that attempts to decrease some predefined 'loss' function which quantifies the magnitude of error predicted by the model with regards to true or observed values. By continuously updating model input parameters, SGD navigates the loss surface via rough approximations of its negative gradient in order to locate some optimal solution on the parameter space that minimizes realization of the loss function for the problem at hand. Inspired by SGD's relative simplicity, momentum-based methods improve convergence by blending past gradients into a weighted averaging scheme. Such an approach effectively reduces variability in SGD's trajectory along the loss surface. Consequently, momentum-based methods tend to offer accelerated convergence rates, reduced oscillations in gradient approximations, and more stable updates to the optimum. In this approach, the standard SGD update rule is:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k, \zeta_k)$$

, where α_k is the learning rate and $\nabla f(x_k, \zeta_k)$ is the gradient of the loss evaluated at the k-th step. SGD minimizes optimization problems via:

$$\min_x F(x) := \mathbb{E}_\zeta[f(x, \zeta)]$$

where $f(x, \zeta)$ is the loss associated with some sample ζ and x are the parameters of the model.

Conversely, a Stochastic Heavy Ball (SHB) momentum-based update can take the general form:

$$d_k = (1 - \beta)\nabla f(x_k, \zeta_k) + \beta d_{k-1}$$

$$x_{k+1} = x_k - \alpha_k d_k$$

where $\beta \in [0, 1]$ is the momentum parameter that specifies how previous gradients affect the current update.

2.2 Quasi-Hyperbolic Momentum (QHM)

As stated above, the QHM update rule offers a more general form of momentum that allows finer control over the balance between fresh gradients and accumulated momentum measurements. The inclusion of the additional parameter ν interpolates between pure gradient descent and the stochastic heavy-ball method. The QHM update rule is defined as:

$$d_k = (1-\beta)\nabla f(x_k, \zeta_k) + \beta d_{k-1}$$

$$x_{k+1} = x_k - \alpha_k[(1-\nu_k)\nabla f(x_k, \zeta_k) + \nu_k d_k]$$

.
When ν=0, QHM corresponds to vanilla SGD while ν=1 corresponds to the SHB method. For ν,β,α constant, $\nu = \beta$, QHM becomes a normalized variant of Newton's Accelerated Gradient (NAG) with an additional coefficient of (1 - β) on the stochastic gradient term. By tuning ν and β, QHM generalizes popular methods like NAG, SHB, and SGD.

2.3 Gaps in Current Understanding

Although convergence under QHM methods is guaranteed theoretically, analyzing the effects of high momentum parameters in practice is less well-understood [2]. For our purpose, the effects of $\beta \to 0$, as $\nu, \beta \to 1$, and the non-linear interaction between β and ν remain unexplored empirically although they are claimed to be true by the authors [2]. This paper aims to explore these gaps in the literature by confirming how QHM behaves asymptotically as the hyperparameters are tuned accordingly.

METHODS AND EXPERIMENTS

3.1 Experimental Setup

To assess the theoretical claims made in the paper, three primary momentum-based methods were implemented:

- QHM: tested with varying parameters β and ν.
- Nesterov's Accelerated Gradient (NAG): a well-regarded technique that incorporates an anticipation of future gradients.
- Stochastic Heavy Ball (SHB): a classic momentum approach known for its simplicity.

We tested the momentum-based methods by utilizing the MNIST [3] and CIFAR-10 [4] datasets for binary image classification tasks. As widely recognized benchmarks, these datasets provided a reliable foundation for testing QHM momentum methods under typical machine learning conditions.

3.2 Metrics for Evaluation

For each algorithm, we recorded the final loss, the convergence rate, stability, and the stationary distribution. The main metrics include:

- Convergence Rate: tracked by the number of iterations needed to reach a predefined loss threshold
- Final Loss: the achieved loss level after convergence to optimum
- Accuracy score: proportion of correctly classified samples
- Stability: how reliably the algorithm converges without divergence, oscillation, or noise

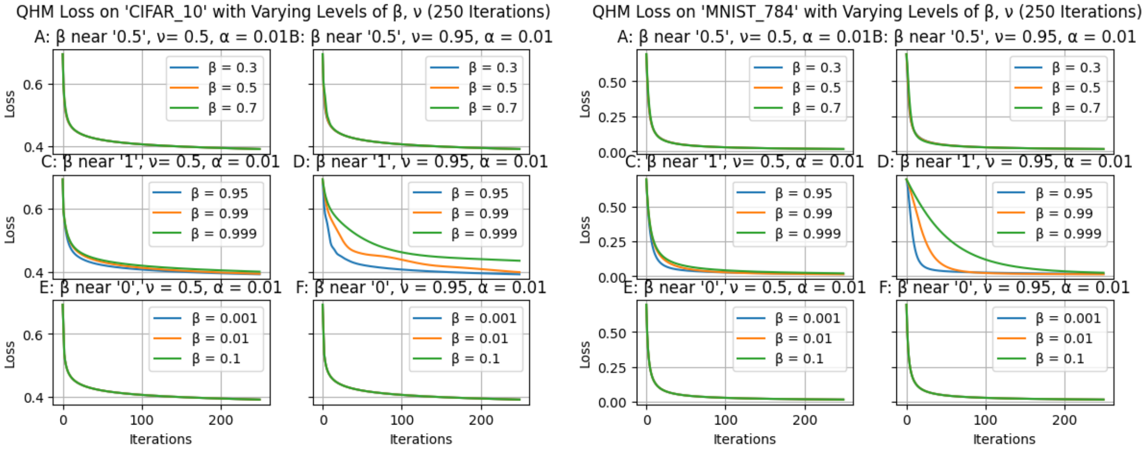
3.3 Experimental Design

We designed a systematic grid search to evaluate a range of parameter settings:

- Momentum parameter β: Values from 0.0 to 1.0 to examine the effects of differing momentum values on convergence. Particularly, we examine the cases for β near zero and 0.5 with arbitrary ν and β near one with ν approaching one
- Interpolation parameter ν: Ranging from 0.0 to 11.0 to study its interaction with β.
- Learning rate α: Values from 0.001 to 0.1, with specific tuning for each momentum method.

Each experiment was repeated multiple times to ensure consistency of results. For each algorithm and configuration, we tuned the learning rate through

Figure 1



grid search and varied β and ν systematically to assess their effects on performance.

RESULTS

4.1 Impact of Increasing β

As expected from the theoretical postulates proposed by the authors, decreasing β towards 0 generally matched or improved QHM's convergence speed and final loss value for all other values of β. However, for β, ν approaching 1, the results were mixed and necessitate an explanation.

Figure 1 depicts the effect of varying β on QHM for arbitrary values of ν. We observe that for β approaching 0, the final loss, rate of convergence, and general stability of QHM appears superior to that for β approaching 1 on both datasets. Nonetheless, all algorithms terminate with nearly identically accuracy scores after 250 iterations. Interestingly, the results are essentially identical to that for β ranging from 0.3 to 0.7, suggesting that the relationship between β and QHM convergence may not be so clear as suggested by the authors.

Figure 2 analyzes QHM convergence for β,$\nu \rightarrow 1$ more closely as the learning rate $\alpha \rightarrow 0$. While decreasing the learning rate generally leads to

Figure 2

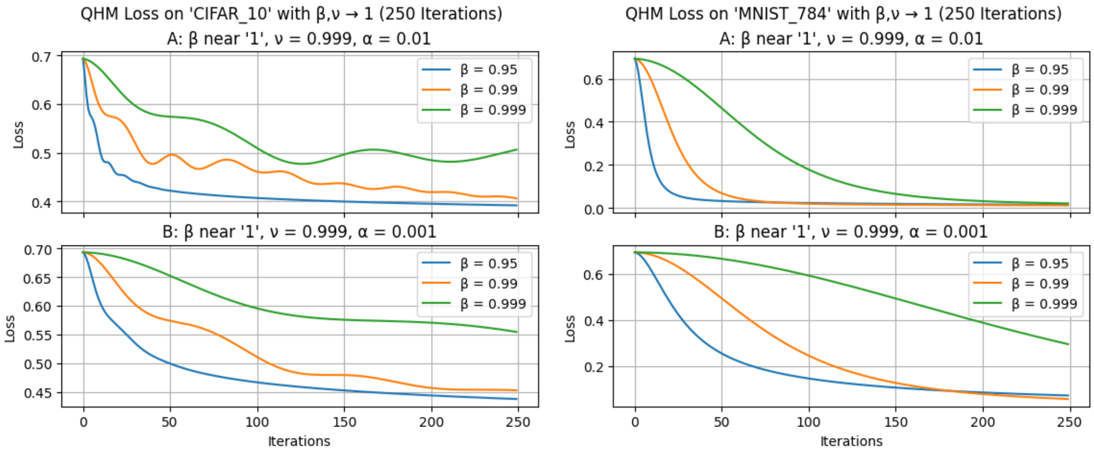
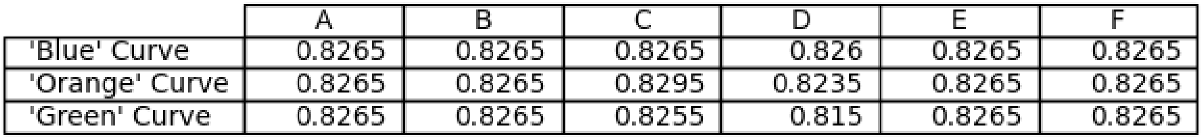
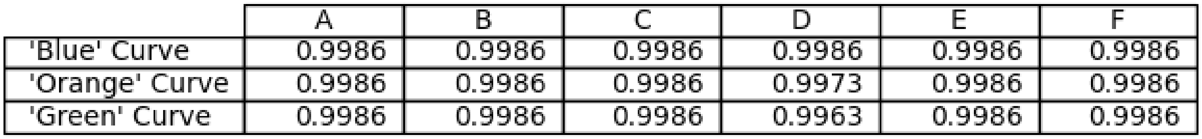


Figure 3: Accuracy Scores for Varying Level of β on 'CIFAR-10','MNIST-784', respectively.



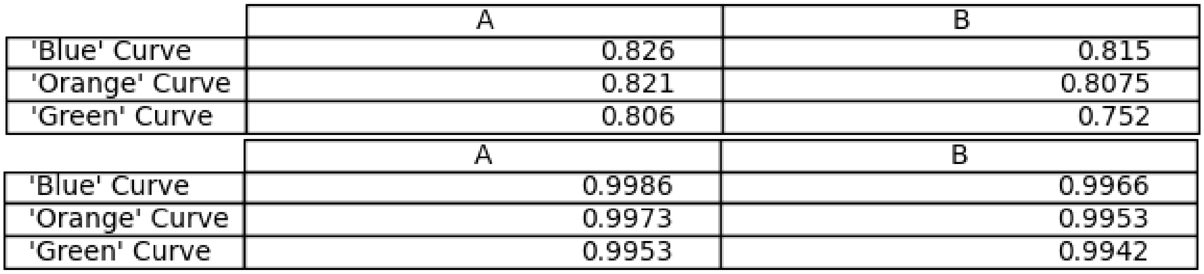
	A	B	C	D	E	F
'Blue' Curve	0.8265	0.8265	0.8265	0.826	0.8265	0.8265
'Orange' Curve	0.8265	0.8265	0.8295	0.8235	0.8265	0.8265
'Green' Curve	0.8265	0.8265	0.8255	0.815	0.8265	0.8265



	A	B	C	D	E	F
'Blue' Curve	0.9986	0.9986	0.9986	0.9986	0.9986	0.9986
'Orange' Curve	0.9986	0.9986	0.9986	0.9973	0.9986	0.9986
'Green' Curve	0.9986	0.9986	0.9986	0.9963	0.9986	0.9986

slower convergence (as expected), stability across the two datasets is far superior for the smaller step size. Such a result is congruent with the authors who claim that QHM does converge for β,ν $\to$ 1 as long as it is slow enough relative to the speed of $\alpha \to 0$. Nonetheless, our experiments across both datasets suggest that QHM convergence under these conditions (β,$\nu$$\to$1 slow enough relative to $\alpha \to 0$) is slower, less stable, and less accurate than QHM under arbitrary β, ν. Accuracy scores are tabulated in *Figure 3* for the first experiment and *Figure 4* for the second.

Figure 4: Accuracy Scores for $\beta\nu \to 1$ on 'CIFAR-10','MNIST-784', respectively.



	A	B
'Blue' Curve	0.826	0.815
'Orange' Curve	0.821	0.8075
'Green' Curve	0.806	0.752

	A	B
'Blue' Curve	0.9986	0.9966
'Orange' Curve	0.9973	0.9953
'Green' Curve	0.9953	0.9942

Figure 5

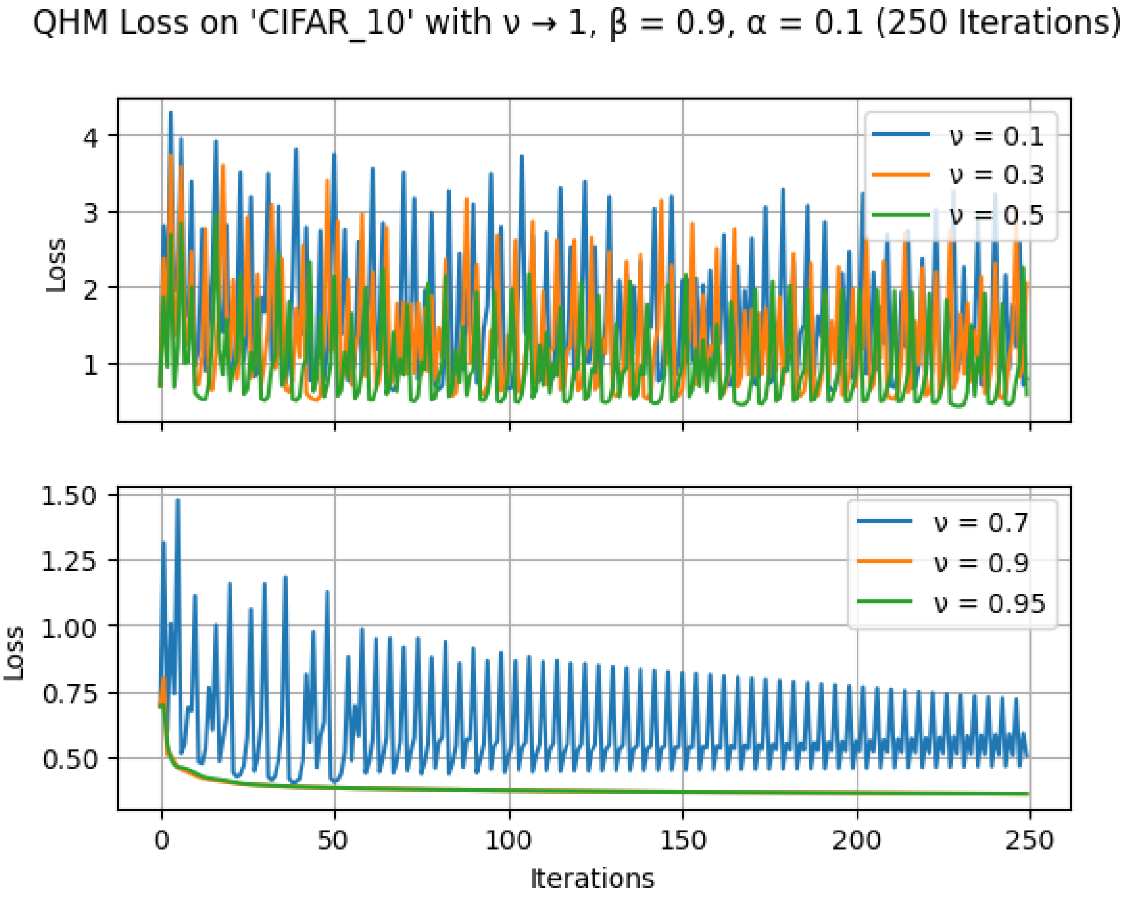
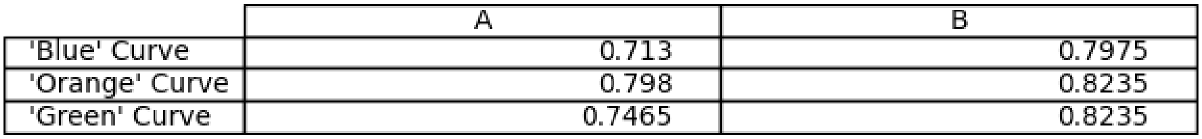


Figure 6: Accuracy Scores for $\nu \to 1$ on 'CIFAR-'10'

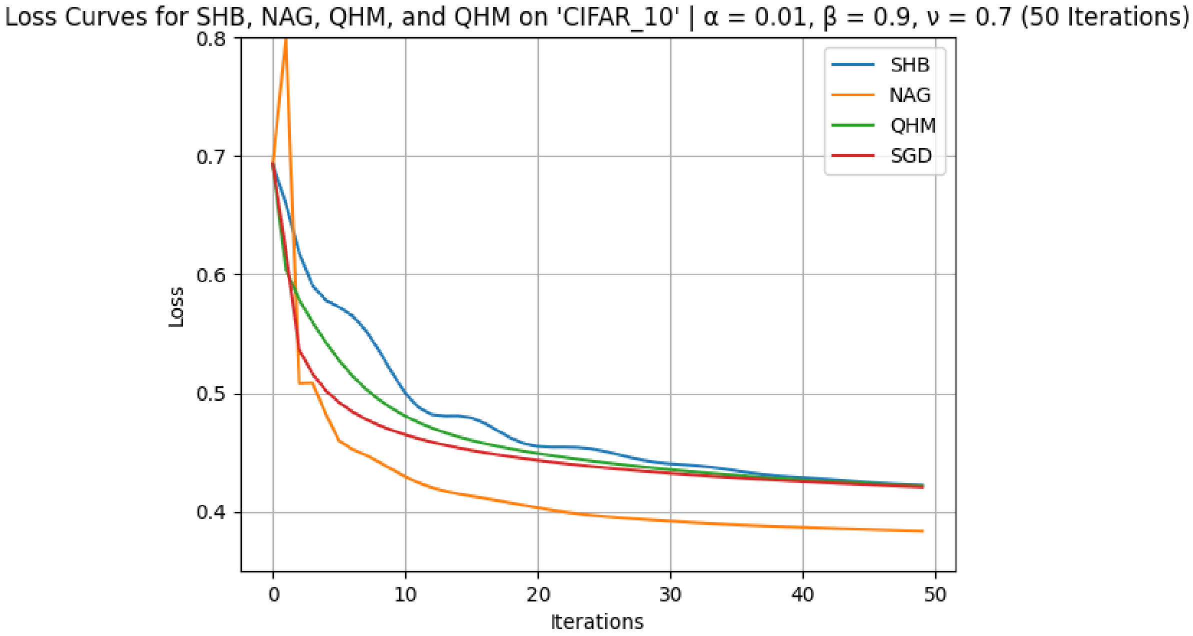


	A	B
'Blue' Curve	0.713	0.7975
'Orange' Curve	0.798	0.8235
'Green' Curve	0.7465	0.8235

4.2 Nonlinear Dependence on ν and β

Results confirmed a non-linear interaction between ν and β, with convergence improvements from increasing β becoming less pronounced at low values of ν. When ν was moderately high, such as 0.9, the gains from increasing β were more substantial but leveled off around $\nu = 0.95$. This interaction emphasizes the need for a balanced configuration of β and ν to achieve optimal performance, as excessively low values of ν lead to significantly more noise and reduced stability but high values of both do not necessarily guarantee additional benefits. A chart of the loss is presented in *Figure 5* and corresponding accuracy scores are tabulated in *Figure 6*.

Figure 7



4.3 Comparative Performance of NAG, QHM, SHB, and SGD

In both experiments, NAG significantly outperformed SHB, general QHM, and SGD in terms of convergence speed and accuracy despite suffering a brief, initial period of instability while traversing the parameter space on 'CIFAR-10'. Of course, such a result is expected due to NAG's "accelerated gradient" component which considers second-order information (i.e. the curvature of loss) to lead to a more direct path towards the optimum [5]. Regardless, SHB, QHM, and SGD all converged at nearly identical rates with accuracy scores just below that of NAG, solidifying previous assumptions that such optimizers are sufficient for learning problems of a certain complexity. It's worth noting that SHB, QHM, and SGD may be more ideal for learning problems in which NAG's propensity for accelerated movements may lead to instability and divergence.

CONCLUSION

Our findings offer a nuanced perspective of the performance and stability of momentum-based gradient methods under the QHM framework. By systematically exploring parameter configurations for both the β-momentum and ν-interpolation parameters via experimentation, we affirmed certain theoretical predictions about key dynamics affecting convergence rates, stability, and accuracy across learning tasks. Interestingly, we also provided evidence that convergence benefits may actually plateau as $\beta,\nu \to 1$, especially at high learning rates, a result that contradicts the findings of the authors. Comparatively, NAG outperformed both QHM, SHB, and SGD in terms of accuracy and speed, underscoring its utility in more complex tasks, albeit with a potential trade-off in stability.

References

[1] Yan Pan and Yuanzhi Li. Toward understanding why adam converges faster than sgd for transformers. 2023. URL https://arxiv.org/abs/2306.00204.

[2] Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. 2019. URL https://arxiv.org/abs/1910.13962.

[3] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[4] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

[5] Gilles Bareilles, Franck Iutzeler, and Jérôme Malick. Newton acceleration on manifolds identified by proximal gradient methods. *Mathematical Programming*, 200(1):37–70, August 2022. ISSN 1436-4646. doi: 10.1007/s10107-022-01873-w. URL http://dx.doi.org/10.1007/s10107-022-01873-w.