We thank the reviewers for their time, effort, and helpful feedback. We will make sure to correct all typos and incorporate the reviewers' suggestions in the final version of the paper. We address individual feedback below.

Reviewer 1:

Yes, "stochastic noise" is the error in the gradient approximations, i.e. $\nabla F(x^k) - g^k$ in the notation of the paper. We will add this definition to the paper.

Bounded noise is indeed a strong assumption, and we are not sure whether the proof can be extended to the more general case where $\mathbf{E}_k[||\xi^k||^2]$ is uniformly bounded. We should note that [25] proves a similar result for SGM with a more general noise condition than bounded noise, and their technique may extend to QHM, but bounded noise greatly simplifies the proof. We will comment on this in the paper. In our opinion, the result showing convergence of QHM when $\nu_k\beta_k \to 1$ is interesting from both theoretical and practical perspectives. From the theoretical side, the result shows that it is possible to always be increasing the amount of momentum (in the limit when $\nu_k\beta_k = 1$, we are not using the fresh gradient information at all!) and still obtain convergence (which is not too surprising, since the objective function is smooth). From the practical point of view, Theorem 5 shows that increasing $\nu_k\beta_k \to 1$ might lead to smaller stationary distribution size, which may give better empirical results. We will include this discussion in the final version of the paper.

Theorem 3 guarantees that there exists some sequence ϵ_k such that the statement holds. We will update the statement of Theorem 3 to include $\mu > 0$ and $\exists\{\epsilon_k\}$ with $\epsilon_k \geq 0$. When $\beta = 1$ and is constant, the algorithm cannot converge, since the gradient information is not used at all. However, as we show in Section 2, it is possible to set $\beta_k \to 1$ in the limit and still obtain convergence. As mentioned above, we found this difference theoretically interesting and will highlight it after Theorem 3.

We had to limit the explanation of Figures 1, 2, and 3 due to space constraints. We would use the additional page allowed in the final version to include more interpretation of these figures, explaining their relationships to Theorems 3, 4, and 5 in more detail.

To improve the presentation of our theoretical results, we will include more lead-in discussion before our theorems (especially Theorem 3) to make them easier to understand. We believe that combining our literature review into its own section (as suggested by Reviewer 2 and addressed below) will also help the reader better contextualize our theorems. Regarding sensitivity, we will expand Section 5 with comments that the presented results depend continuously on the algorithm parameters.

Reviewer 2:

Since we presented a range of results covering different aspects of momentum methods (asymptotic convergence, stability, and stationary distributions), and each of these areas has its own rich literature, we put a brief review of the relevant works at the beginning of each section. To improve readability, we will combine these references in a separate section to give a narrative that ties the different areas together.

We cite Mandt et al. [16] for some empirical justification of our Assumption A3 (in the beginning of Section 3) and for proving an equivalent result to our Theorem 4 in the case of stochastic heavy ball (in the beginning of Section 4). We can make the existing citation after Theorem 4 more prominent.

By saying that "we use quadratic functions for ease of analysis" we did not mean to imply that it is straightforward to prove our results for a more general class of functions, although the basic principles remain similar. We will change the wording of this phrase in the final version of the paper to make it more specific. The results in Theorems 4 and 5 may hold approximately near a local minimum for nonconvex but sufficiently smooth functions, when the learning rate is sufficiently small. In that case, the function is intuitively almost quadratic locally, since the higher-order terms in the Taylor expansion around the optimum can be neglected.

For most of our experiments, the code is very minimal. To make it easier to reproduce our results, we will release the code for all of the experiments performed in this paper.

Reviewer 3:

We agree with the reviewer that a more general analysis is necessary to better understand the properties of momentum methods in the wild, especially for the nonconvex problems faced in deep learning practice. While our results are based on some restrictive assumptions, we hope they still provide valuable practical insights, and we believe they could serve as an important foundation for any future work extending the unified analysis of momentum methods to more general classes of functions.