Unified Momentum Dynamics in Stochastic Gradient Optimization

Chris Junchi Li$^\diamond$

Department of Electrical Engineering and Computer Sciences$^\diamond$
University of California, Berkeley

October 9, 2024

Abstract

Momentum-based methods have become foundational in stochastic optimization, particularly for training deep neural networks. Despite their widespread use, there remains a significant gap in understanding the interplay between various momentum formulations, including Stochastic Heavy Ball (SHB) and Nesterov Accelerated Gradient (NAG), and their impact on convergence and performance. This paper explores a broad class of momentum algorithms, generalizing the Quasi-Hyperbolic Momentum (QHM) framework, and presents theoretical insights into their convergence behavior under both diminishing and constant learning rates. We analyze asymptotic convergence, local stability, and stationary distributions, offering practical guidelines for parameter tuning in deep learning. Empirical evidence is provided to demonstrate how these insights can lead to more efficient optimization strategies.

Keywords: Stochastic Gradient Descent, Momentum Methods, Quasi-Hyperbolic Momentum, Convergence Analysis, Optimization

1 Introduction

Stochastic gradient descent (SGD) and its variants have been the cornerstone of optimization in machine learning, particularly for large-scale nonconvex problems such as training deep neural networks. Over the years, numerous modifications of the basic SGD algorithm have been proposed, with momentum methods standing out as a key improvement for accelerating convergence. Momentum techniques, such as the Stochastic Heavy Ball (SHB) method and Nesterov Accelerated Gradient (NAG), are widely recognized for their ability to smooth the optimization path and avoid oscillations in poorly conditioned regions of the parameter space.

However, the precise role of momentum in improving both convergence rates and final performance remains only partially understood. Recent empirical studies have yielded mixed conclusions: while some highlight the critical role of momentum for fast convergence, others claim that certain parameter regimes show little to no practical improvement. This raises important questions about how momentum should be parametrized in different stages of optimization.

In this paper, we contribute to bridging this theoretical gap by providing a unified framework that generalizes many well-known momentum methods, including SHB, NAG, and Quasi-Hyperbolic Momentum (QHM). Our analysis focuses on the asymptotic convergence of these methods under diminishing learning rates, as well as their local stability and behavior when constant parameters are used. Additionally, we explore the stationary distribution of the parameters, which provides key insights into how the final performance of these methods can be optimized.

Backgrounds Stochastic gradient methods have become extremely popular in machine learning for solving stochastic optimization problems of the form

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ F(x) \triangleq \mathbf{E}_\zeta \big[f(x, \zeta) \big] \tag{1}$$

where ζ is a random variable representing data sampled from some (unknown) probability distribution, $x \in \mathbb{R}^n$ represents the parameters of a machine learning model (e.g., the weight matrices in a neural network), and f is a loss function associated with the model parameters and any sample ζ. Many variants of the stochastic gradient methods can be written in the form of

$$x^{k+1} = x^k - \alpha_k d^k \tag{2}$$

where d^k is a (stochastic) search direction and $\alpha_k > 0$ is the step size or learning rate. The classical stochastic gradient descent (SGD) [Robbins and Monro(1951)] method uses $d^k = \nabla_x f(x^k, \zeta^k)$, where ζ^k is a random sample collected at step k. For the ease of notation, we use g^k to denote $\nabla_x f(x^k, \zeta^k)$ throughout this paper.

There is a vast literature on modifications of SGD that aim to improve its theoretical and empirical performance. The most common such modification is the addition of a *momentum* term, which sets the search direction d^k as the combination of the current stochastic gradient g^k and past search directions. For example, the stochastic variant of Polyak's heavy ball method [Polyak(1964)] uses

$$d^k = g^k + \beta_k d^{k-1} \tag{3}$$

where $\beta_k \in [0, 1)$. We call the combination of (2) and (3) the Stochastic Heavy Ball (SHB) method. [Gupal and Bazhenov(1972)] studied a "normalized" version of SHB, where

$$d^k = (1 - \beta_k) g^k + \beta_k d^{k-1} \tag{4}$$

In the context of modern deep learning, [Sutskever et al.(2013)] proposed to use a stochastic variant of Nesterov's accelerated gradient (NAG) method, where

$$d^k = \nabla_x f \big(x^k - \alpha_k \beta_k d^{k-1}, \zeta^k \big) + \beta_k d^{k-1} \tag{5}$$

The number of variations on momentum has kept growing in recent years; see, e.g., Synthesized Nesterov Variants (SNV) [Lessard et al.(2016)], Triple Momentum [Van Scoy et al.(2017)], Robust Momentum [Cyrus et al.(2018)], PID Control-based methods [An et al.(2018)], Accelerated SGD (Acc-SGD) [Kidambi et al.(2018)], and Quasi-Hyperbolic Momentum (QHM) [Ma and Yarats(2019)].

Despite various empirical successes reported for these different methods, there is a lack of clear understanding of how the different forms of momentum and their associated parameters affect convergence properties of the algorithms and other performance measures, such as final loss value. For example, [Sutskever et al.(2013)] show that momentum is critical to obtaining good performance in deep learning. But using different parametrizations, [Ma and Yarats(2019)] claim that momentum may have little practical effect. In order to clear up this confusion, several recent works [Yang et al.(2016), An et al.(2018), Ma and Yarats(2019), see, e.g.,] have aimed to develop and analyze general frameworks that capture many different momentum methods as special cases.

In this paper, we focus on a class of algorithms captured by the general form of QHM [Ma and Yarats(2019)]:

$$\begin{aligned} d^k &= (1 - \beta_k) g^k + \beta_k d^{k-1}, \\ x^{k+1} &= x^k - \alpha_k \Big[(1 - \nu_k) g^k + \nu_k d^k \Big] \end{aligned} \tag{6}$$

where the parameter $\nu_k \in [0, 1]$ interpolates between SGD ($\nu_k = 0$) and (normalized) SHB ($\nu_k = 1$). When the parameters α_k, β_k and ν_k are held constant (thus the subscript k can be omitted) and $\nu = \beta$, it recovers a normalized variant of NAG with an additional coefficient $1 - \beta_k$ on the stochastic gradient term in (5) (see Appendix A). In addition, [Ma and Yarats(2019)] show that different settings of α_k, β_k and ν_k recover other variants such as AccSGD [Kidambi et al.(2018)], Robust Momentum [Cyrus et al.(2018)], and Triple Momentum [Van Scoy et al.(2017)]. They also show that it is equivalent to SNV [Lessard et al.(2016)] and special cases of PID Control (either PI or PD) [An et al.(2018)]. However, there is little theoretical analysis of QHM in general. In this paper, we take advantage of its general formulation to derive a unified set of analytic results that help us better understand the role of momentum in stochastic gradient methods.

1.1 Contributions and outline

Our theoretical results on the QHM model (6) cover three different aspects: asymptotic convergence with probability one, stability region and local convergence rates, and characterizations of the stationary distribution of $\{x^k\}$ under constant parameters α, β, and ν. Specifically:

- In Section 3, we show that for minimizing smooth nonconvex functions, QHM converges almost surely as $\beta_k \to 0$ for arbitrary values of ν_k. And more surprisingly, we show that QHM converges as $\nu_k \beta_k \to 1$ (which requires both $\nu_k \to 1$ and $\beta_k \to 1$) as long as $\nu_k \beta_k \to 1$ slow enough, as compared with the speed of $\alpha_k \to 0$.

- In Section 4, we consider local convergence behaviors of QHM for fixed parameters α, β, and ν. In particular, we derive joint conditions on (α, β, ν) that ensure local stability (or convergence when there is no stochastic noise in the gradient approximations) of the algorithm near a strict local minimum. We also characterize the local convergence rate within the stability region.

- In Section 5, we investigate the stationary distribution of $\{x^k\}$ generated by the QHM dynamics around a local minimum (using a simple quadratic model with noise). We derive the dependence of the stationary variance on (α, β, ν) up to the second-order Taylor expansion in α. These results reveal interesting effects of β and ν that cannot be seen from first-order expansions.

Our asymptotic convergence results in Section 3 give strong guarantees for the convergence of QHM with diminishing learning rates under different regimes ($\beta_k \to 0$ and $\beta_k \to 1$). However, as with most asymptotic results, they provide limited guidance on how to set the parameters in practice for fast convergence. Our results in Sections 4 and 5 complement the asymptotic results by providing principled guidelines for tuning these parameters. For example, one of the most effective schemes used in deep learning practice is called "constant and drop", where constant parameters (α, β, ν) are used to train the model for a long period until it reaches a stationary state and then the learning rate α is dropped by a constant factor for refined training. Each stage of the constant-and-drop scheme runs variants of QHM with constant parameters, and their choices dictate the overall performance of the algorithm. In Section 6, by combining our results in Sections 4 and 5, we obtain new and, in some cases, counter-intuitive insight into how to set these parameters in practice.

2 Related work

Asymptotic convergence   There exist many classical results concerning the asymptotic convergence of the stochastic gradient methods; see, e.g., [Wasan(1969), Polyak(1987), Kushner and Yin(2003)]

and references therein. For the classical SGD method without momentum, i.e., (2) with $d^k = g^k$, a well-known general condition for asymptotic convergence is $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. In general, we will always need $\alpha_k \to 0$ to counteract the effect of noise. But interestingly, the conditions on β_k are much less restricted. For normalized SHB, [Polyak(1977)] and [Kaniovski(1983)] studied its asymptotic convergence properties in the regime of $\alpha_k \to 0$ and $\beta_k \to 0$, while [Gupal and Bazhenov(1972)] investigated asymptotic convergence in the regime of $\alpha_k \to 0$ and $\beta_k \to 1$, both for *convex* optimization problems. More recently, [Gadat et al.(2018)] extended asymptotic convergence analysis for the normalized SHB update to smooth nonconvex functions for $\beta_k \to 1$. In this work we generalize the classical SGD and SHB results to the case of QHM for smooth nonconvex functions.

Local convergence rate   The stability region and local convergence rate of the deterministic gradient descent and heavy ball algorithms were established by Boris Polyak for the case of convex functions near a strict twice-differentiable local minimum [Polyak(1963), Polyak(1964)]. For this class of functions heavy ball method is optimal in terms of the local convergence rate [Nemirovsky and Yudin(1983)]. However, it might fail to converge globally for the general strongly convex twice-differentiable functions [Lessard et al.(2016)] and is no longer optimal for the class of smooth convex functions. For the latter case, Nesterov's accelerated gradient was shown to attain the optimal *global* convergence rate [Nesterov(1983), Nesterov(2004)]. In this paper we extend the results of [Polyak(1964)] on local convergence to the more general QHM algorithm.

Stationary analysis   The limit behavior analysis of SGD algorithms with momentum and constant step size was used in various applications. [Pflug(1983), Yaida(2018), Lang et al.(2019)] establish sufficient conditions on detecting whether iterates reach stationarity and use them in combination with statistical tests to automatically change learning rate during training. [Freidlin and Wentzell(1998), Dieuleveut et al.(2017)] prove many properties of limiting behavior of SGD with constant step size by using tools from Markov chain theory. Our results are most closely related to the work of [Mandt et al.(2017)] who use stationary analysis of SGD with momentum to perform approximate Bayesian inference. In fact, our Theorem 4 extends their results to the case of QHM and our Theorem 5 establishes more precise relations (to the second order in α), revealing interesting dependence on the parameters β and ν which cannot be seen from the first order equations.

3   Asymptotic convergence

In this section, we generalize the classical asymptotic results to provide conditions under which QHM converges almost surely to a stationary point for smooth *nonconvex* functions. Throughout this section, "a.s." refers to "almost surely". We need to make the following assumptions.

Assumption 1. *The following conditions hold for F defined in* (1) *and the stochastic gradient oracle:*

  1. F is differentiable and ∇F is Lipschitz continuous, i.e., there is a constant L such that

$$\|\nabla F(x) - \nabla F(y)\| \le L\|x - y\|, \qquad x, y \in \mathbb{R}^n$$

  2. F is bounded below and $\|\nabla F(x)\|$ is bounded above, i.e., there exist F_* and G such that

$$F(x) \ge F_*, \qquad \|\nabla F(x)\| \le G, \qquad x \in \mathbb{R}^n$$

4

3. For $k = 0, 1, 2, \ldots$, the stochastic gradient $g^k = \nabla F(x^k) + \xi^k$, where the random noise ξ^k satisfies

$$\mathbf{E}_k[\xi^k] = 0, \qquad \mathbf{E}_k\big[\|\xi^k\|^2\big] \leq C \quad a.s.$$

where $\mathbf{E}_k[\cdot]$ denotes expectation conditioned on $\{x^0, g^0, \ldots, x^{k-1}, g^{k-1}, x^k\}$, and C is a constant.

Note that Assumption 1.3 allows the distribution of ξ^k to depend on x^k, and we simply require the second moment to be conditionally bounded uniformly in k. The assumption $\|\nabla F(x)\| \leq G$ can be removed if we assume a bounded domain for x. However, this will complicate the proof by requiring special treatment (e.g., using the machinery of gradient mapping [Nesterov(2013)]) when $\{x^k\}$ converges to the boundary of the domain. Here we assume this condition to simplify the analysis.

By convergence to a stationary point, we mean that the sequence $\{x^k\}$ satisfies the condition

$$\liminf_{k\to\infty} \|\nabla F(x^k)\| = 0 \quad a.s \tag{7}$$

Intuitively, as $\beta_k \to 0$, regardless of ν_k, the QHM dynamics become more like SGD, so there should be no issue with convergence. The following theorem, which generalizes the analysis technique of [Ruszczyński and Syski(1984)] to QHM, shows formally that this is indeed the case:

Theorem 1. *Let F satisfy Assumption 1. Additionally, assume $0 \leq \nu_k \leq 1$ and the sequences $\{\alpha_k\}$ and $\{\beta_k\}$ satisfy the following conditions:*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty, \qquad \lim_{k\to\infty} \beta_k = 0, \qquad \overline{\beta} \triangleq \sup_k \beta_k < 1$$

Then the sequence $\{x^k\}$ generated by the QHM algorithm (6) satisfies (7). Moreover, we have

$$\limsup_{k\to\infty} F(x^k) = \limsup_{k\to\infty,\ \|\nabla F(x^k)\|\to 0} F(x^k) \quad a.s \tag{8}$$

More surprisingly, however, one can actually send $\nu_k \beta_k \to 1$ as long as $\nu_k \beta_k \to 1$ slow enough, although we require a stronger condition on the noise ξ. We extend the technique of [Gupal and Bazhenov(1972)] to show asymptotic convergence of QHM for minimizing smooth non-convex functions.

Theorem 2. *Let F satisfy assumption 1, and additionally assume that $\|\xi^k\|^2 < C$ almost surely, i.e., the noise ξ is a.s. bounded. Let the sequences $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\nu_k\}$ satisfy the following conditions:*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty}(1 - \nu_k\beta_k)^2 < \infty, \qquad \sum_{k=0}^{\infty} \frac{\alpha_k^2}{1 - \nu_k\beta_k} < \infty, \qquad \lim_{k\to\infty} \beta_k = 1$$

Then the sequence $\{x^k\}$ generated by Algorithm (6) satisfies (7).

The conditions in Theorem 2 can be satisfied by, for example, taking $\alpha_k = k^{-\omega}$ and $(1 - \nu_k\beta_k) = k^{-c}$ for $\frac{1+c}{2} < \omega \leq 1$ and $\frac{1}{2} < c < 1$. We should note that, even though setting $\nu_k\beta_k \to 1$ is somewhat unusual in practice, we think the result of Theorem 2 is interesting from both theoretical and practical points of view. From the theoretical side, this 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 for smooth functions. From the practical point of view, our Theorem 5 in Section 5 shows that for a fixed α, increasing $\nu_k \beta_k$ might lead to smaller stationary distribution size, which may give better empirical results.

Also, note that when $\nu_k = \beta_k$, Theorems 1 and 2 give asymptotic convergence guarantees for the common practical variant of NAG, which have not appeared in the literature before. However, we should mention that the bounded noise assumption of Theorem 2 (i.e. $\|\xi^k\|^2 < C$ a.s.) is quite restrictive. In fact, [Ruszczyński and Syski(1983)] prove a similar result for SGM with a more general noise condition, and their technique may extend to QHM, but bounded noise greatly simplifies the derivations. We provide the proofs of Theorems 1 and 2 in Appendix B.

The results in this section indicate that both $\beta_k \to 0$ and $\nu_k \beta_k \to 1$ are admissible from the perspective of asymptotic convergence. However, they give limited guidance on how to choose momentum parameters in practice, where non-asymptotic behaviors are of main concern. In the next two sections, we study local convergence and stationary behaviors of QHM with constant learning rate and momentum parameters; our analysis provides new insights that could be very useful in practice.

4   Stability region and local convergence rate

Let the sequence $\{x^k\}$ be generated by the QHM algorithm (6) with constant parameters $\alpha_k = \alpha$, $\beta_k = \beta$ and $\nu_k = \nu$. In this case, x^k does not converge to any local minimum in the asymptotic sense, but its distribution may converge to a stationary distribution around a local minimum. Since the objective function F is smooth, we can approximate F around a strict local minimum x^* by a convex quadratic function. Since $\nabla F(x^*) = 0$, we have

$$F(x) \approx F(x^*) + \tfrac{1}{2}(x - x^*)^T \nabla^2 F(x^*)(x - x^*)$$

where the Hessian $\nabla^2 F(x^*)$ is positive definite. Therefore, for the ease of analysis, we focus on convex quadratic functions of the form $F(x) = (1/2)(x-x^*)^T A(x-x^*)$, where A is positive definite (and we can set $x^* = 0$ without loss of generality). In addition, we assume

$$g^k = \nabla F(x^k) + \xi^k = A(x - x^*) + \xi^k \tag{9}$$

where the noise ξ^k satisfies Assumption 1.3 and in addition, ξ^k is independent of x^k for all $k \geq 0$. [Mandt et al.(2017)] observe that this independence assumption often holds approximately when the dynamics of SHB are approaching stationarity around a local minimum.

Under the above assumptions, the behaviors of QHM can be described by a linear dynamical system driven by i.i.d. noise. More specifically, let $z^k = [d^{k-1}; x^k - x^*] \in \mathbb{R}^{2n}$ be an augmented state vector, then the dynamics of (6) can be written as (see Appendix E for details)

$$z^{k+1} = T z^k + S \xi^k \tag{10}$$

where T and S are functions of (α, β, ν) and A:

$$T = \begin{bmatrix} \beta I & (1-\beta)A \\ -\alpha\nu\beta I & I - \alpha(1-\nu\beta)A \end{bmatrix}, \qquad S = \begin{bmatrix} (1-\beta)I \\ -\alpha(1-\nu\beta)I \end{bmatrix} \tag{11}$$

It is well-known that the linear system (10) is stable if and only if the spectral radius of T, denoted by $\rho(T)$, is less than 1. When $\rho(T) < 1$, the dynamics of (10) is the superposition of two components:

- A deterministic part described by the dynamics $z^{k+1} = Tz^k$ with initial condition $z^0 = [0; x^0]$ (we always take $d^{-1} = 0$). This part asymptotically decays to zero.

- An auto-regressive stochastic process (10) driven by $\{\xi^k\}$ with zero initial condition $z^0 = [0; 0]$. Roughly speaking, $\rho(T)$ determines how fast the dynamics converge from an arbitrary initial point x^0 to the stationary distribution, while properties of the stationary distribution (such as its variance and auto-correlations) depends on the full spectrum of the matrix T as well as S. Both aspects have important implications for the practical performance of QHM on stochastic optimization problems. Often there are trade-offs that we have to make in choosing the parameters α, β and ν to balance the transient convergence behavior and stationary distribution properties.

In the rest of this section, we focus on the deterministic dynamics $z^{k+1} = Tz^k$ to derive the conditions on (α, β, ν) that ensure $\rho(T) < 1$ and characterize the convergence rate. Let $\lambda_i(A)$ for $i = 1, \ldots, n$ denote the eigenvalues of A (they are all real and positive). In addition, we define

$$\mu = \min_{i=1,\ldots,n} \lambda_i(A), \qquad L = \max_{i=1,\ldots,n} \lambda_i(A), \qquad \kappa = L/\mu$$

where κ is the condition number. The local convergence rate for strictly convex quadratic functions is well studied for the case of gradient descent ($\nu = 0$) and heavy ball ($\nu = 1$) [Polyak(1964)]. In fact, heavy ball achieves the best possible convergence rate of $(\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$[Nesterov(2004)]. Thus, it is immediately clear that the optimal convergence rate of QHM will be the same and will be achieved with $\nu = 1$. However, there are no results in the literature characterizing how the optimal rate or optimal parameters change as a function of ν. Our next result establishes the convergence region and dependence of the convergence rate on the parameters α, β, and ν. We present the result for quadratic functions, but it can be generalized to any L-smooth and μ-strongly convex functions, assuming the initial point x^0 is close enough to the optimal point x_* (see Theorem 6 in Appendix C).

Theorem 3. *Let's denote $\theta = \{\alpha, \beta, \nu\}^1$. For any function $F(x) = x^T A x + b^T x + c$ that satisfies $0 < \mu \leq \lambda_i(A) \leq L$ for all $i = 1, \ldots, n$ and any x^0, $\exists \{\epsilon_k\}$, with $\epsilon_k \geq 0$, such that the deterministic QHM algorithm $z^{k+1} = Tz^k$ satisfies*

$$\left\| x^k - x_* \right\| \leq (R(\theta, \mu, L) + \epsilon_k)^k \left\| x^0 - x_* \right\|$$

where $x_* = \arg\min_x F(x)$, $\lim_{k \to \infty} \epsilon_k = 0$ and $R(\theta, \mu, L) = \rho(T)$, which can be characterized as

$$R(\theta, \mu, L) = \max\{r(\theta, \mu), r(\theta, L)\}, \quad \text{where}$$

$$r(\theta, \lambda) = \begin{cases} 0.5\left(\sqrt{C_1(\lambda)^2 - 4C_2(\lambda)} + C_1(\lambda)\right) & \text{if } C_1(\lambda) \geq 0, C_1(\lambda)^2 - 4C_2(\lambda) \geq 0, \\ 0.5\left(\sqrt{C_1(\lambda)^2 - 4C_2(\lambda)} - C_1(\lambda)\right) & \text{if } C_1(\lambda) < 0, C_1(\lambda)^2 - 4C_2(\lambda) \geq 0, \\ \sqrt{C_2(\lambda)} & \text{if } C_1(\lambda)^2 - 4C_2(\lambda) < 0, \end{cases} \quad (12)$$
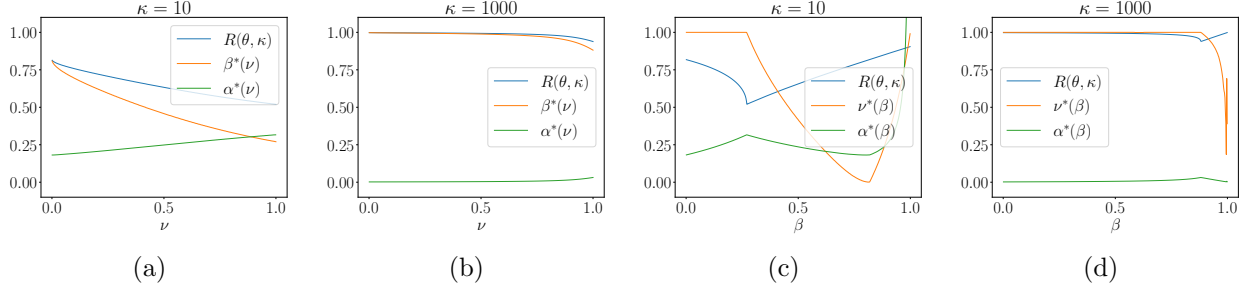
$$C_1(\lambda, \theta) = 1 - \alpha\lambda + \alpha\lambda\nu\beta + \beta,$$
$$C_2(\lambda, \theta) = \beta(1 - \alpha\lambda + \alpha\lambda\nu)$$

To ensure $R(\theta, \mu, L) < 1$, the parameters α, β, ν must satisfy the following constraints:

$$0 < \alpha < \frac{2(1+\beta)}{L(1 + \beta(1 - 2\nu))}, \qquad 0 \leq \beta < 1, \qquad 0 \leq \nu \leq 1 \quad (13)$$

In addition, the optimal rate depends only on κ: $\min_\theta R(\theta, \mu, L)$ is a function of only κ.

7

Figure 1. Plots (a), (b) show the dependence of the optimal α, β and convergence rate as a function of ν. We can see that both rate and optimal β are decreasing functions of ν. Plots (c), (d) show the dependence of optimal α, ν and rate on β. We can see that there are three phases in which the dependence is quite different. Also note that in all presented cases, changing ν required changing α in the same way (they increase and decrease together).

The conditions in (13) characterize the stability region of QHM. Note that when $\nu = 0$ we have the classical result for gradient descent: $\alpha < 2/L$; when $\nu = 1$, the condition matches that of the normalized heavy ball: $\alpha < 2(1 + \beta)/(L(1 - \beta))$.

The equations (12) define the convergence rate for any fixed values of the parameters α, β, ν. While it does not give a simple analytic form, it allows us to conduct easy numerical investigations. To gain more intuition into the effect that momentum parameters ν and β have on the convergence rate, we study how the optimal ν changes as a function of β and vice versa. To find the optimal parameters and rate, we solve the corresponding optimization problem numerically (using the procedure described in Appendix D). For each pair $\{\beta, \nu\}$ we set α to the optimal value in order to remove its effect. These plots are presented in Figure 1.

A natural way to think about the interplay between parameters α, β and ν is in terms of the total "amount of momentum". Intuitively, it should be controlled by the product of $\nu \times \beta$. This intuition helps explain Figure 1 (a), (b), which show the dependence of the optimal β as a function of ν for different values of κ. We can see that for bigger values of ν we need to use smaller values of β, since increasing each one of them increases the "amount of momentum" in QHM. However, the same intuition fails when considering ν as a function of β (and β is big enough), as shown in Figure 1 (c), (d). In this case there are 3 regimes of different behavior. In the first regime, since β is small, the amount of momentum is not enough for the problem and thus the optimal ν is always 1. In this phase we also need to increase α when increasing β (it is typical to use larger learning rate when the momentum coefficient is bigger). The second phase begins when we reach the optimal value of β (rate is minimal) and, after that, the amount of momentum becomes too big and we need to decrease ν and α. However, somewhat surprisingly, there is a third phase, when β becomes big enough we need to start increasing ν and α again. Thus we can see that it's not just the product of $\nu\beta$ that governs the behavior of QHM, but a more complicated function.

Finally, based on our analytic and numerical investigations, we conjecture that the optimal convergence rate is a *monotonically decreasing* function of ν (if α and β are chosen optimally for each ν). While we can't prove this statement[2], we verify this conjecture numerically in Appendix D. The code of all of our experiments is available at `https://github.com/Kipok/understanding-momentum`.

---

[1]We drop the dependence of some functions on θ for brevity.

[2]In fact, we hypothesise that $R^*(\nu, \kappa)$ might not have analytical formula, since it is possible to show that the optimization problem over α and β is equivalent to the system of highly non-linear equations.
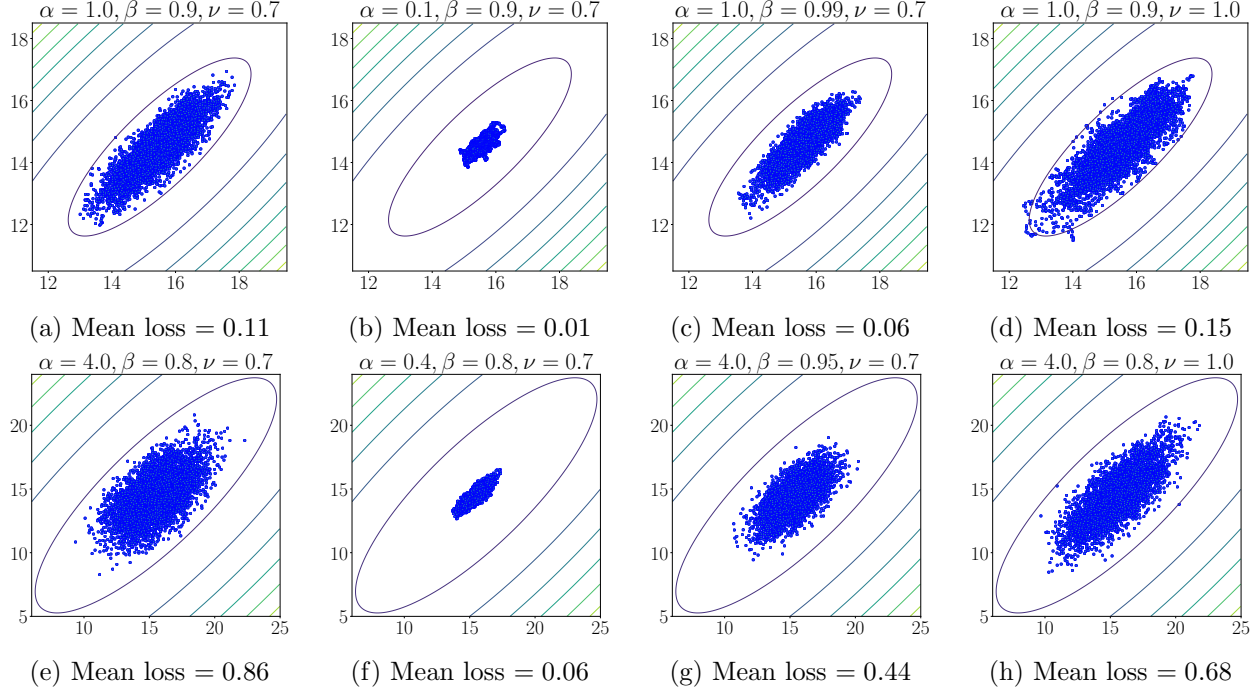
Figure 2. Changes in the shape and size of stationary distribution changes with respect to α, β, and ν on a 2-dimensional quadratic problem. Each picture shows the last 5000 iterates of QHM on a contour plot. The first picture of each row is a reference and other pictures should be compared to it. The second pictures show how the stationary distribution changes when we decrease α. The third and fourth show the dependence on β and ν, respectively. We can see that as expected, moving $\alpha \to 0$ and $\beta \to 1$ always decreases the achievable loss. However, the dependence on ν is more complicated, and for some values of α and β increasing ν increases the loss (top row), while for other values the dependence is reversed (bottom row). Note the scale change between top and bottom plots.

5   Stationary analysis

In this section, we study the stationary behavior of QHM with constant parameters α, β and ν. Again we only consider quadratic functions for the same reasons as outlined in the beginning of Section 4. In other words, we focus on the linear dynamics of (10) driven by the noise ξ^k as $k \to \infty$ (where the deterministic part depending on x^0 dies out). Under the assumptions of Section 4 we have the following result on the covariance matrix defined as $\Sigma_x \triangleq \lim_{k \to \infty} \mathbf{E}\big[x^k(x^k)^T\big]$.

Theorem 4. *Suppose $F(x) = \frac{1}{2}x^T A x$, where A is symmetric positive definite matrix. The stochastic gradients satisfy $g^k = \nabla F(x^k) + \xi$, where ξ is a random vector independent of x^k with zero mean $\mathbf{E}[\xi] = 0$ and covariance matrix $\mathbf{E}\big[\xi\xi^T\big] = \Sigma_\xi$. Also, suppose the parameters α, β, ν satisfy (13). Then the QHM algorithm (6), equivalently (10) in this case, converges to a stationary distribution satisfying*

$$A\Sigma_x + \Sigma_x A = \alpha A\Sigma_\xi + O(\alpha^2) \tag{14}$$

When $\nu = 1$, this result matches the known formula for the stationary distribution of unnormalized SHB [Mandt et al.(2017)] with reparametrization of $\alpha \to \alpha/(1-\beta)$. Note that Theorem 4 shows that for the normalized version of the algorithm, the stationary distribution's covariance does not depend on β (or ν) to the first order in α. In order to explore such dependence, we need to expand the dependence on α to the second order. In that case, we are not able to obtain a matrix
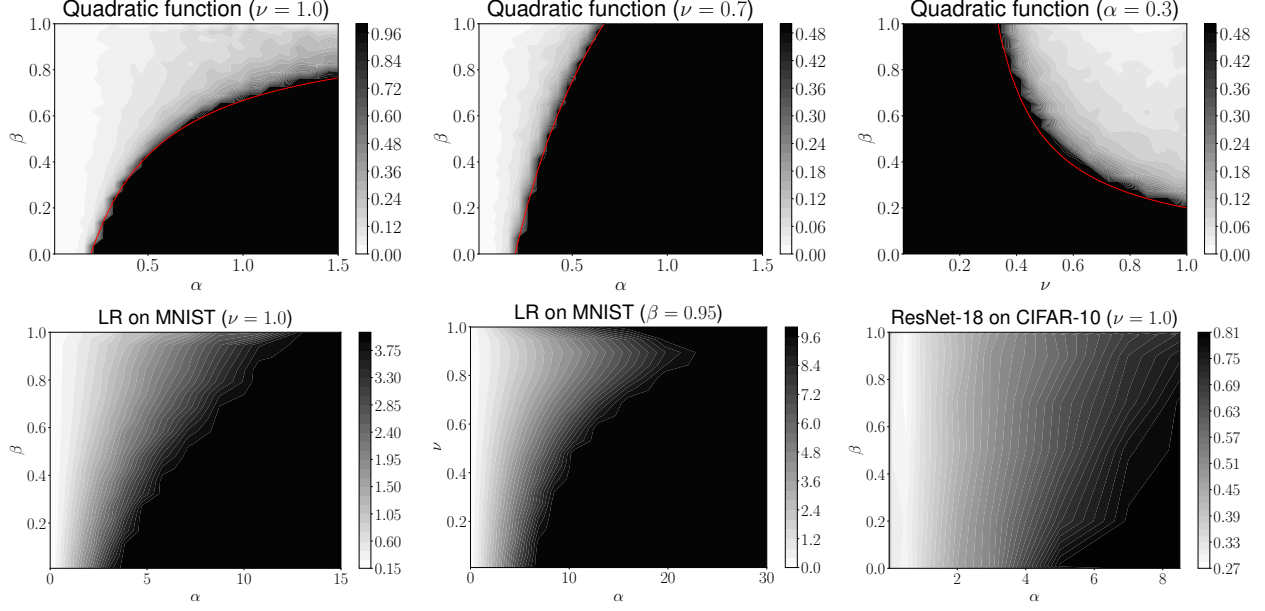
Figure 3. These pictures show dependence of the average final loss (depicted with color: whiter is smaller) on the parameters of QHM algorithm for different problems. The top row shows results for a synthetic 2-dimensional quadratic problem, where all the assumptions of Theorem 5 are satisfied. The red curve indicates the boundary of convergence region (algorithm diverges below it). In this case, we start the algorithm directly at the optimal value to measure the size of stationary distribution and ignore convergence rate. We can see that as predicted by theory, smaller α and bigger β make the final loss smaller. The bottom row shows results of the same experiments repeated for logistic regression on MNIST and ResNet-18 on the CIFAR-10 dataset. We can see that while the assumptions of Theorem 5 are no longer valid, QHM still shows similar qualitative behavior.

equation, but can get the following relation for $\mathbf{tr}(A\Sigma_x)$.

Theorem 5. *Under the conditions of Theorem 4, we have*

$$\mathbf{tr}(A\Sigma_x) = \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\left(1 + \frac{2\nu\beta}{1-\beta}\left[\frac{2\nu\beta}{1+\beta} - 1\right]\right)\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3) \tag{15}$$

We note that $\mathbf{tr}(A\Sigma_x)$ is twice the mean value of $F(x)$ when the dynamics have reached stationarity, so the right-hand side of (15) is approximately the "achievable loss" given the values of α, β and ν. It is interesting to consider several special cases:

- $\nu = 0$ (SGD): $\mathbf{tr}(A\Sigma_x) = \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3)$.

- $\nu = 1$ (SHB): $\mathbf{tr}(A\Sigma_x) = \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\frac{1-\beta}{1+\beta}\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3)$.

- $\nu = \beta$ (NAG): $\mathbf{tr}(A\Sigma_x) = \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\left(1 - \frac{2\beta^2(1+2\beta)}{1+\beta}\right)\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3)$.

From the expressions for SHB and NAG, it might be beneficial to move β to 1 during training in order to make the achievable loss smaller. While moving β to 1 is somewhat counter-intuitive, we proved in Section 3 that QHM still converges asymptotically in this regime, assuming ν also goes to 1 and $\nu\beta$ converges to 1 "slower" than α converges to 0. However, since we only consider Taylor expansion in α, there is no guarantee that the approximation remains accurate when ν and β converge to 1 (see Appendix G for evaluation of this approximation error). In order to precisely investigate the dependence on β and ν, it is necessary to further extend our results by considering

10
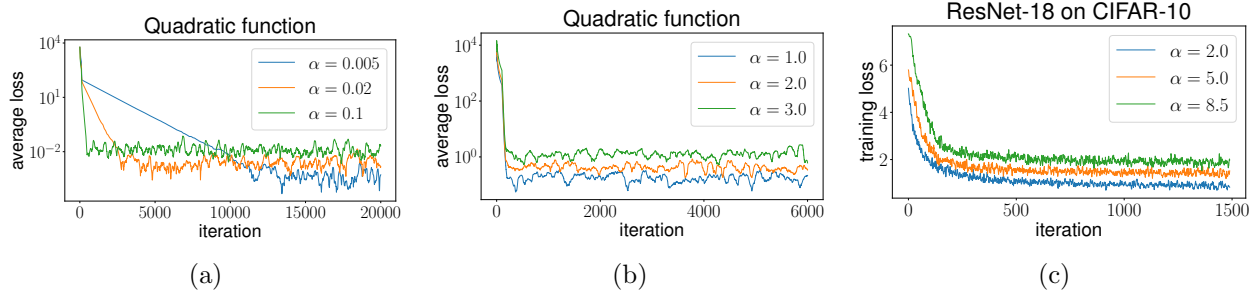
Figure 4. (a) This plot shows a trade-off between stationary distribution size (final loss) and convergence rate on a simple 2-dimensional quadratic problem. Algorithms that converge faster, typically will converge to a higher final loss. (b) This plot illustrates the regime where there is no trade-off between stationary distribution size and convergence rate. Larger values of α don't change the convergence rate, while making final loss significantly higher. To make plots (a) and (b) smoother, we plot the average value of the loss for each 100 iterations on y-axis. (c) This plot shows that the same behavior can also be observed in training deep neural networks. For all plots $\beta = 0.9, \nu = 1.0$. All of the presented results depend continuously on the algorithm's parameters (e.g. the transition between behaviours shown in (a) and (b) is smooth).

Taylor expansion with respect to them as well, especially in terms of $1 - \beta$. We leave this for future work.

Figure 2 shows a visualization of the QHM stationary distribution on a 2-dimensional quadratic problem. We can see that our prediction about the dependence on α and β holds in this case. However, the dependence on ν is more complicated: the top and bottom rows of Figure 2 show opposite behavior. Comparing this experiment with our analysis of the convergence rate (Figure 1) we can see another confirmation that for big values of β, increasing ν can, in a sense, decrease the "amount of momentum" in the system. Next, we evaluate the average final loss for a large grid of parameters α, β and ν on three problems: a 2-dimensional quadratic function (where all of our assumptions are satisfied), logistic regression on the MNIST [LeCun(1998)] dataset (where the quadratic assumption is approximately satisfied, but gradient noise comes from mini-batches) and ResNet-18 [He et al.(2016)] on CIFAR-10 [Krizhevsky and Hinton(2009)] (where all of our assumptions are likely violated). Figure 3 shows the results of this experiment. We can indeed see that $\beta \to 1$ and $\alpha \to 0$ make the final loss smaller in all cases. The dependence on ν is less clear, but we can see that for large values of β it is approximately quadratic, with a minimum at some $\nu < 1$. Thus from this point of view $\nu \neq 1$ helps when β is big enough, which might be one of the reasons for the empirical success of the QHM algorithm. Notice that the empirical dependence on ν is qualitatively the same as predicted by formula (15), but with optimal value shifted closer to 1. See Appendix F for details.

6 Some practical implications and guidelines

In this section, we present some practical implications and guidelines for setting learning rate and momentum parameters in practical machine learning applications. In particular, we consider the question of how to set the optimal parameters in each stage of the popular constant-and-drop scheme for deep learning. We argue that in order to answer this question, it is necessary to consider both convergence rate and stationary distribution perspectives. There is typically a trade-off between obtaining a fast rate and a small stationary distribution. You can see an illustration of this trade-off in Figure 4 (a). Interestingly, by combining stationary analysis of Section 5 and results for the

convergence rate (3), we can find certain regimes of parameters α, β, and ν where the final loss and the convergence speed do not compete with each other.

One of the most important of these regimes happens in the case of the SHB algorithm ($\nu = 1$). In that case, we can see that when $C_1^2(l) - C_2^2(l) \leq 0, l \in \{\mu, L\}$, the convergence rate equals $\sqrt{\beta}$ and does not depend on α. Thus, as long as this inequality is satisfied, we can set α as small as possible and it would not harm the convergence rate, but will decrease the size of stationary distribution. To get the best possible convergence rate, we, in fact, have to set α and β in such a way that this inequality will turn into equality and thus there will be only a single value of α that could be used. However, as long as β is not exactly at the optimal value, there is going to be some freedom in choosing α and it should be used to decrease the size of stationary distribution. From this point of view, the optimal value of $\alpha = \left(1 - \sqrt{\beta}\right) / \left(\mu\left(1 + \sqrt{\beta}\right)\right)$, which will be smaller then the largest possible α for convergence as long as $\kappa > 2$ and β is set close to 1 (see proof of Theorem 3 for more details). This guideline contradicts some typical advice to set α as big as possible while algorithm still converges[3]. The refined guideline for the constant-and-drop scheme would be to set α as small as possible until the convergence noticeably slows down. You can see an illustration of this behavior on a simple quadratic problem (Figure 4 (b)), as well as for ResNet-18 on CIFAR-10 (Figure 4 (c)). Such regimes of no trade-off can be identified for β and ν as well.

7 Conclusion

Using the general formulation of Quasi-Hyperbolic Momentum (QHM), we have derived a unified set of analytic results that provide deeper insights into the role of momentum in stochastic gradient methods. Our analysis addresses several critical aspects, including asymptotic convergence, stability regions, local convergence rates, and the behavior of stationary distributions. These insights highlight the key trade-offs involved in tuning learning rates and momentum parameters to optimize performance in stochastic optimization.

It is important to note that the practical guidelines we have developed are focused primarily on minimizing the *training* loss. There is evidence that different heuristics and parameter settings may be necessary to improve generalization performance in machine learning, but this remains beyond the scope of the current paper.

References

[An et al.(2018)] Wangpeng An, Haoqian Wang, Qingyun Sun, Jun Xu, Qionghai Dai, and Lei Zhang. A pid controller approach for stochastic optimization of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8522–8531, 2018.

[Bengio et al.(2013)] Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. Advances in optimizing recurrent networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8624–8628. IEEE, 2013.

[Cyrus et al.(2018)] Saman Cyrus, Bin Hu, Bryan Van Scoy, and Laurent Lessard. A robust accelerated optimization algorithm for strongly convex functions. In *2018 Annual American Control Conference (ACC)*, pages 1376–1381. IEEE, 2018.

---

[3]For example [Goh(2017)] says "So set β as close to 1 as you can, and then find the highest α which still converges. Being at the knife's edge of divergence, like in gradient descent, is a good place to be."

[Dieuleveut et al.(2017)] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *arXiv preprint arXiv:1707.06386*, 2017.

[Ermoliev(1969)] Yu M Ermoliev. On the stochastic quasi-gradient method and stochastic quasi-feyer sequences. *Kibernetika*, 2:72–83, 1969.

[Freidlin and Wentzell(1998)] Mark Iosifovich Freidlin and Alexander D Wentzell. Random perturbations. In *Random perturbations of dynamical systems*, pages 15–43. Springer, 1998.

[Gadat et al.(2018)] Sébastien Gadat, Fabien Panloup, Sofiane Saadane, et al. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018.

[Goh(2017)] Gabriel Goh. Why momentum really works. *Distill*, 2017. doi: 10.23915/distill.00006. URL `http://distill.pub/2017/momentum`.

[Gupal and Bazhenov(1972)] A. M. Gupal and L. T. Bazhenov. A stochastic analog of the conjugate gradient method. *Cybernetics*, 8(1):138–140, 1972.

[He et al.(2016)] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Kaniovski(1983)] Yu. M. Kaniovski. Behaviour in the limit of iterations of the stochastic two-step method. *USSR Computational Mathematics and Mathematical Physics*, 23(1):8–13, 1983.

[Kidambi et al.(2018)] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[Krizhevsky and Hinton(2009)] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[Kushner and Yin(2003)] Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition, 2003.

[Lang et al.(2019)] Hunter Lang, Pengchuan Zhang, and Lin Xiao. Statistical adaptive stochastic approximation. 2019.

[LeCun(1998)] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[Lessard et al.(2016)] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[Ma and Yarats(2019)] Jerry Ma and Denis Yarats. Quasi-hyperbolic momentum and adam for deep learning. In *International Conference on Learning Representations*, 2019.

[Mandt et al.(2017)] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

[Meyer(1972)] Paul-André Meyer. *Martingales and stochastic integrals I*, volume 284 of *Lecture notes in mathematics*. Springer-Verlag, 1972.

[Nemirovsky and Yudin(1983)] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[Nesterov(1983)] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. In *Soviet Math. Dokl*, volume 27, 1983.

[Nesterov(2004)] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004.

[Nesterov(2013)] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[Pflug(1983)] Georg Ch. Pflug. On the determination of the step size in stochastic quasigradient methods. Collaborative Paper CP-83-025, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria, 1983.

[Polyak(1964)] Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[Polyak(1977)] Boris T. Polyak. Comparison of the rates of convergence of one-step and multi-step optimization algorithms in the presence of noise. *Engineering Cybernetics*, 15:6–10, 1977.

[Polyak(1987)] Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.

[Polyak(1963)] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[Recht(2010)] Benjamin Recht. Cs726-lyapunov analysis and the heavy ball method. 2010.

[Robbins and Monro(1951)] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[Ruszczyński and Syski(1983)] Andrzej Ruszczyński and Wojciech Syski. Stochastic approximation method with gradient averaging for unconstrained problems. *IEEE Transactions on Automatic Control*, 28(12): 1097–1105, 1983.

[Ruszczyński and Syski(1984)] Andrzej Ruszczyński and Wojciech Syski. Stochastic approximation algorithm with gradient averaging and on-line stepsize rules. In J. Gertler and L. Keviczky, editors, *Proceedings of 9th IFAC World Congress*, pages 1023–1027, Budapest, Hungary, 1984.

[Sutskever et al.(2013)] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[Tange et al.(2011)] Ole Tange et al. Gnu parallel-the command-line power tool. *The USENIX Magazine*, 36(1):42–47, 2011.

[Van Scoy et al.(2017)] Bryan Van Scoy, Randy A Freeman, and Kevin M Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2017.

[Wasan(1969)] M. T. Wasan. *Stochastic Approximation*. Cambridge University Press, 1969.

[Williams(1991)] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[Yaida(2018)] Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. *arXiv preprint arXiv:1810.00004*, 2018.

[Yang et al.(2016)] Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

A   From NAG to QHM

In this appendix we will mention exact steps needed to come from the original NAG formulation to the formulation assumed by the QHM algorithm. We refer the reader to the ([Sutskever et al.(2013)] Appendix A.1) for the derivation of NAG as the following momentum method [4]

$$d_k = \beta_{k-1} d_{k-1} - \alpha_{k-1} \nabla f(x_{k-1} + \beta_{k-1} d_{k-1})$$
$$x_k = x_{k-1} + d_k$$

Next, we will move the learning rate out of the momentum into the iterates update:

$$d_k = \beta_{k-1} d_{k-1} + \nabla f(x_{k-1} - \alpha_{k-1} \beta_{k-1} d_{k-1})$$
$$x_k = x_{k-1} - \alpha_{k-1} d_k$$

When α_k and β_k are constant, the two methods produce the same sequence of iterates x_k if d_0 is initialized at 0. To make the notation more similar to the QHM algorithm, let's move all indices (except for d_k) up by 1:

$$d_k = \beta_k d_{k-1} + \nabla f(x_k - \alpha_k \beta_k d_{k-1})$$
$$x_{k+1} = x_k - \alpha_k d_k$$

This again does not change the algorithm. Now, let's normalize the momentum update by $1 - \beta_k$:

$$d_k = \beta_k d_{k-1} + (1 - \beta_k) \nabla f(x_k - \alpha_k \beta_k d_{k-1})$$
$$x_{k+1} = x_k - \alpha_k d_k$$

This version is equivalent to the unnormalized by re-scaling $\alpha \to \alpha/(1 - \beta)$ for constant parameters[5]. Finally, following [Bengio et al.(2013)] we need to make a change of variables $y_k = x_k - \alpha_k \beta_k d_{k-1}$ and additionally assume that $\beta_k = \beta$ is constant:

$$d_k = \beta d_{k-1} + (1 - \beta) \nabla f(y_k)$$
$$y_{k+1} = x_{k+1} - \alpha_k \beta d_k = x_k - \alpha_k d_k - \alpha_k \beta d_k = y_k + \alpha_k \beta d_{k-1} - \alpha_k d_k - \alpha_k \beta d_k$$
$$= y_k + \alpha_k (d_k - (1 - \beta) \nabla f(y_k)) - \alpha_k d_k - \alpha_k \beta d_k$$
$$= y_k - \alpha_k [(1 - \beta) \nabla f(y_k) + \beta d_k]$$

Renaming y_k back to x_k and replacing $\nabla f(y_k)$ with stochastic gradient if necessary we obtain the exact formula used in QHM update.

Overall, assuming $d_0 = 0$ and β_k is constant, the QHM version of NAG is indeed equivalent (up to a change of variable) to the original NAG with re-scaling of $\alpha \to \alpha/(1 - \beta)$. However, if β_k is changing from iteration to iteration, the two algorithms are no longer equivalent.

B   Asymptotic Convergence Proofs

In this section we prove Theorems 1 and 2. For simplicity, we assume throughout that α_k, ν_k and β_k are nonrandom.

Proof of Theorem 1. Here we generalize the meta-analysis of [Ruszczyński and Syski(1984)] to include ν_k.

---

[4]Note that we change the notation to be consistent with the notation of QHM.
[5]In fact, for non-constant β_k the algorithms are no longer equivalent.

Algorithm. We consider the following variant of QHM:

$$d^k = g^k + i_k \beta_k d^{k-1} \tag{16}$$

$$b^k = (1 - \nu_k)g^k + \nu_k d^k \tag{17}$$

$$x^{k+1} = x^k - \alpha_k b^k \tag{18}$$

where ν_k is in $[0, 1]$ and i_k is a binary switch introduced to handle unbounded noise. Specifically, for some constant $\rho > 0$, we let

$$i_k = \begin{cases} 1 & \text{if } \|d^{k-1}\| \le \rho, \\ 0 & \text{if } \|d^{k-1}\| > \rho \end{cases} \tag{19}$$

Conditions repeated for convenience. Let F satisfy Assumption 1. Additionally, assume the sequences $\{\alpha_k\}$, $\{\beta_k\}$ and $\{\nu_k\}$ satisfy the following conditions:

$$\sum_{k=0}^{\infty} \alpha_k = \infty \tag{20}$$

$$\sum_{k=0}^{\infty} \alpha_k^2 < \infty \tag{21}$$

$$\lim_{k \to \infty} \beta_k = 0 \tag{22}$$

$$\overline{\beta} \triangleq \sup_k \beta_k < 1 \tag{23}$$

$$0 \le \nu_k \le 1$$

Then the dynamics (6) satisfy (7) and (8).

Remark. A direct consequence of Assumption 1.1 is that, for all $x, y \in \mathbb{R}^n$,

$$\left| F(y) - F(x) - \langle \nabla F(x), y - x \rangle \right| \le \frac{L}{2} \|y - x\|^2 \tag{24}$$

which we will use often in the convergence analysis.

We first prove the following key lemma.

Lemma 1. *Suppose Assumption 1, (20), and (23) hold, and for every $k \ge 0$,*

$$F(x^{k+1}) \le F(x^k) - \alpha_k(1 - \nu_k \beta_k)\|\nabla F(x^k)\| \left(\|\nabla F(x^k)\| - S_k \right) + W_k \tag{25}$$

where $\{S_k\}$ and $\{W_k\}$ are sequences of scalar random variables satisfying and

$$\lim_{k \to \infty} S_k = 0 \quad a.s. \tag{26}$$

$$\sum_{k=0}^{\infty} W_k < \infty \quad a.s \tag{27}$$

Then we have

$$\liminf_{k \to \infty} \|\nabla F(x^k)\| = 0 \quad a.s \tag{28}$$

and

$$\limsup_{k \to \infty} F(x^k) = \limsup_{k \to \infty, \ \|\nabla F(x^k)\| \to 0} F(x^k) \quad a.s \tag{29}$$

That is, (8) and (7) both hold.

16

Proof. For the first part, suppose for a contradiction that $\liminf_{k\to\infty} ||\nabla F(x^k)|| > 0$. So there exists $\epsilon > 0$ and k_0 such that for all $k \geq k_0$, $||\nabla F(x^k)|| \geq \epsilon$. By (26), there exists $k_1 \geq k_0$ such that $S_k \leq \epsilon/2$ for all $k \geq k_1$. Then from (25) we obtain:

$$F(x^{k+1}) \leq F(x^k) - \frac{\epsilon}{2}\alpha_k(1 - \nu_k\beta_k)||\nabla F(x^k)|| + W_k$$

Summing over k from k_1 to ∞ and using Assumption 1.2, we get that:

$$\frac{\epsilon}{2}\sum_{k=k_1}^{\infty} \alpha_k(1 - \nu_k\beta_k)||\nabla F(x^k)|| \leq F(x^{k_1}) - F^* + \sum_{k=k_1}^{\infty} W_k$$

The right-hand-side is finite by (27). But since $||\nabla F(x^k)|| \geq \epsilon$ for all $k \geq k_1$ and $\beta_k \leq \overline{\beta} < 1$ and $0 \leq \nu_k \leq 1$, we have:

$$\frac{\epsilon^2}{2}(1 - \overline{\beta})\sum_{k=k_1}^{\infty} \alpha_k \leq F(x^{k_1}) - F^* + \sum_{k=k_1}^{\infty} W_k$$

This implies $\sum_{k=k_1}^{\infty} \alpha_k < \infty$, which contradicts (20). So we must have (7), i.e. $\liminf_k ||\nabla F(x^k)|| = 0$.

To prove (8), we consider two cases. First, assume there exists k_0 such that $||\nabla F(x^k)|| \geq S_k$ for all $k \geq k_0$. Then by (7), there exists a subsequence $\mathcal{K} \subset \mathbb{N}$ such that

$$\lim_{k\in\mathcal{K}, k\to\infty} ||\nabla F(x^k)|| = 0$$

For every l, define the index $k(l) = \max\{k \in \mathcal{K} : k < l\}$. Since $\mathcal{K}$ is infinite, $k(l) \to \infty$ as $l \to \infty$. Then for sufficiently large l, i.e., when $k(l) \geq k_1$, (25) becomes

$$F(x^l) \leq F(x^{k(l)}) + \sum_{i=k(l)}^{l-1} W_i$$

As $l \to \infty$, because of (27) and $k(l) \to \infty$, we get $\sum_{i=k(l)}^{l-1} W_i \to 0$, so

$$\limsup_{l\to\infty} F(x^l) \leq \limsup_{l\to\infty} F(x^k(l)) \leq \limsup_{k\in\mathcal{K}, k\to\infty} F(x^k) \tag{30}$$

Since the reverse inequality is trivial, we obtain (8).

In the second case, we have $||\nabla F(x^k)|| < S_k$ fulfilled infinitely often. In that case, for each l define the index $k(l) = \max\{k : k < l \text{ and } ||\nabla F(x^k)|| < S_k\}$. As before, $k(l) \to \infty$ as $l \to \infty$. Furthermore, (26) implies $||\nabla F(x^{k(l)})|| \to 0$ as $l \to \infty$. Therefore, there exists $\mathcal{K} \subset \mathbb{N}$ with $\{k(l)\}_l \subset \mathcal{K}$ and $\lim_{k\in\mathcal{K}, k\to\infty} ||\nabla F(x^k)|| = 0$. In this case, we obtain from (25) that

$$F(x^l) \leq F(x^{k(l)}) + \alpha_{k(l)}(1 - \nu_{k(l)}\beta_{k(l)})||\nabla F(x^{k(l)})||S_{k(l)} + \sum_{i=k(l)}^{l-1} W_i$$

Because $\alpha_k < \overline{\alpha} < \infty$ for all k, ν_k, β_k are in $[0, 1]$, and $||\nabla F(x^{k(l)})|| \to 0$, the latter two terms in the above inequality converge to zero as $l \to \infty$. So we obtain (30) again. This concludes the proof of the lemma. $\square$

All that remains now is to use the smoothness inequality (24) to identify the sequences S_k and W_k for the dynamics of the modified algorithm (16)-(18), and prove (26) and (27).

From the update formula (18) and using $g^k = \nabla F(x^k) + \xi^k$, we obtain

$$\Delta x^{k+1} = x^{k+1} - x^k = -\alpha_k b^k$$
$$= -\alpha_k b^k + \alpha_k (1 - \nu_k \beta_k)\big(g^k - \nabla F(x^k) - \xi^k\big)$$
$$= -\alpha_k (1 - \nu_k \beta_k)\nabla F(x^k) - \alpha_k \big(b^k - (1 - \nu_k \beta_k)g^k\big) - \alpha_k (1 - \nu_k \beta_k)\xi^k$$

By the smoothness assumption 1.1, we have

$$F(x^{k+1}) \le F(x^k) + \langle \nabla F(x^k), \Delta x^{k+1}\rangle + \frac{L}{2}\big\|\Delta x^{k+1}\big\|^2$$
$$= F(x^k) - \alpha_k(1 - \nu_k \beta_k)\big\|\nabla F(x^k)\big\|^2 - \alpha_k \langle \nabla F(x^k), b^k - (1 - \nu_k \beta_k)g^k\rangle$$
$$\qquad - \alpha_k(1 - \nu_k \beta_k)\langle \nabla F(x^k), \xi^k\rangle + \frac{L}{2}\big\|\Delta x^{k+1}\big\|^2$$
$$\le F(x^k) - \alpha_k(1 - \nu_k \beta_k)\big\|\nabla F(x^k)\big\|^2 + \alpha_k\big\|\nabla F(x^k)\big\| \cdot \big\|b^k - (1 - \nu_k \beta_k)g^k\big\|$$
$$\qquad - \alpha_k(1 - \nu_k \beta_k)\langle \nabla F(x^k), \xi^k\rangle + \frac{L}{2}\big\|\Delta x^{k+1}\big\|^2$$
$$= F(x^k) - \alpha_k(1 - \nu_k \beta_k)\big\|\nabla F(x^k)\big\|\left(\big\|\nabla F(x^k)\big\| - \frac{\big\|b^k - (1 - \nu_k \beta_k)g^k\big\|}{(1 - \nu_k \beta_k)}\right)$$
$$\qquad - \alpha_k(1 - \nu_k \beta_k)\langle \nabla F(x^k), \xi^k\rangle + \frac{L}{2}\big\|\Delta x^{k+1}\big\|^2$$

Comparing with (25), we define

$$S_k = \frac{\big\|b^k - (1 - \nu_k \beta_k)g^k\big\|}{1 - \nu_k \beta_k} \tag{31}$$

$$W_k = -\alpha_k(1 - \nu_k \beta_k)\langle \nabla F(x^k), \xi^k\rangle + \frac{L}{2}\big\|\Delta x^{k+1}\big\|^2 \tag{32}$$

First we show $S_k \to 0$. From the update formula, we have

$$b^k = (1 - \nu_k)g^k + \nu_k\Big((1 - \beta_k)g^k + i_k\beta_k d^{k-1}\Big) = (1 - \nu_k \beta_k)g^k + i_k \nu_k \beta_k d^{k-1}$$

Then because $i_k\|d^{k-1}\| \le \rho$, $\beta_k \to 0$, and $\sup_k \beta_k = \overline{\beta} < 1$, we have

$$\lim_{k\to\infty} S_k = \lim_{k\to\infty}\frac{\|b^k - (1 - \nu_k\beta_k)g^k\|}{1 - \nu_k\beta_k} = \lim_{k\to\infty}\frac{i_k\nu_k\beta_k\|d^{k-1}\|}{1 - \nu_k\beta_k} \le \lim_{k\to\infty}\frac{\nu_k\beta_k\rho}{1 - \nu_k\overline{\beta}} \le \lim_{k\to\infty}\frac{\beta_k\rho}{1 - \overline{\beta}} = 0$$

Because $S_k \ge 0$, $\lim_k S_k = 0$.

Now we show that W_k is summable almost surely. To begin, we need to show that $\|\Delta x^{k+1}\|^2$ is summable, for which we need the following lemma:

Lemma 2. *There is a random variable C, constant in k, such that $\mathbf{E}_k[\|b^k\|^2] \le C$ for all k almost surely.*

Proof. To see this, observe that

$$\|b^k\| = \Big\|(1 - \nu_k)g^k + \nu_k\Big((1 - \beta_k)g^k + i_k\beta_k d^{k-1}\Big)\Big\| = \|g^k - \nu_k g^k + \nu_k g^k - \nu_k\beta_k g^k + i_k\nu_k\beta_k d^{k-1}\|$$
$$= \|(1 - \nu_k\beta_k)g^k + i_k\nu_k\beta_k d^{k-1}\| \le (1 - \nu_k\beta_k)\|g^k\| + i_k\nu_k\beta_k\|d^{k-1}\| \le (1 - \nu_k\beta_k)\|g^k\| + \nu_k\beta_k\rho$$

18

where in the last inequality we used $i_k||d^{k-1}|| \le \rho$. Then

$$\mathbf{E}_k[||b^k||^2] \le (1 - \nu_k\beta_k)^2\mathbf{E}_k[||g^k||^2] + (1 - \nu_k\beta_k)\nu_k\beta_k\rho\mathbf{E}_k[||g^k||] + \nu_k^2\beta_k^2\rho^2$$

By assumption 1.2 and 1.3, the first and second conditional moments $\mathbf{E}_k[||g^k||]$ and $\mathbf{E}_k[||g^k||^2]$ are both bounded uniformly in k. Then because ν_k and β_k are in $[0, 1]$ and ρ is constant in k, we can put

$$\mathbf{E}_k[||b^k||^2] \le (1 - \nu_k\beta_k)^2 C' + (1 - \nu_k\beta_k)\nu_k\beta_k\rho C'' + \nu_k^2\beta_k^2\rho^2 \le C$$

which is what we wanted. Note that C could be a random variable (depending on ω), but this bound holds almost surely. $\square$

Lemma 3. $\sum_{k=1}^{\infty} ||\Delta x^{k+1}||^2 < \infty$ *almost surely*

Proof. We will use the following useful proposition (known as Levy's sharpening of Borel-Cantelli Lemma, see e.g. [Meyer(1972), Chapter 1, Theorem 21]):

Proposition 1. *Let $\{b_k\}$ be a sequence of positive, integrable random variables, and let $a_i = \mathbf{E}[b_i|\mathcal{F}_i]$, where $\mathcal{F}_i = \{b_0, \dots b_{i-1}\}$. Then defining the partial sums $B_k = \sum_{i=1}^k b_i$, $A_k = \sum_{i=1}^k a_i$,*

$$\lim_{k\to\infty} A_k < \infty \ a.s. \implies \lim_{k\to\infty} B_k < \infty \ a.s.$$

So to prove $\sum_{k=1}^{\infty} ||\Delta x^{k+1}||^2 < \infty$, we only need to prove $\sum_{k=1}^{\infty} \mathbf{E}_k[||\Delta x^{k+1}||^2] < \infty$. To see this, observe that

$$\sum_{k=1}^{\infty} \mathbf{E}_k[||\Delta x^{k+1}||^2] = \sum_{k=1}^{\infty} \alpha_k^2 \mathbf{E}_k[||b^k||^2]$$

so because $\mathbf{E}_k[||b^k||^2] \le C$ a.s.,

$$\sum_{k=1}^{\infty} \mathbf{E}_k[||\Delta x^{k+1}||^2] \le C \sum_{k=1}^{\infty} \alpha_k^2 < \infty \ a.s.$$

where we used (21). Applying the proposition finishes the lemma. $\square$

The last term remaining in W_k is $-\alpha_k(1 - \nu_k\beta_k)\langle\nabla F(x^k), \xi^k\rangle$. We show that

$$M_k = \sum_{i=0}^k \alpha_i(1 - \nu_i\beta_i)\langle\nabla F(x^i), \xi^i\rangle$$

is a convergent martingale. First, note that $\mathbf{E}_i[\alpha_i(1-\nu_i\beta_i)\langle\nabla F(x^i), \xi^i\rangle] = 0$, so $E_k[M_k] = M_{k-1}$, and M_k is a martingale. Now we show that $\sup_k \mathbf{E}[M_k^2]$ is bounded, which will imply a.s. convergence by Doob's forward convergence theorem [Williams(1991), Section 11.5]. Indeed, $\mathbf{E}[M_k^2] = \mathbf{E}[M_0^2] + \sum_{i=1}^k \mathbf{E}[(M_i - M_{i-1})^2]$ [Williams(1991), Section 12.1].

$$\mathbf{E}[M_0^2] = \mathbf{E}[\langle\nabla F(x^0), \alpha_0(1-\nu_0\beta_0)\xi^0\rangle^2] \le \mathbf{E}[||\nabla F(x^0)||^2 \cdot \alpha_0^2(1-\nu_0\beta_0)^2||\xi^0||^2] \le G\mathbf{E}[\alpha_0^2(1-\nu_0\beta_0)^2||\xi^0||^2]$$

Because ξ^0 depends only on x_0, we can upper bound this expectation with some constant C by using Assumption 1.3. Then we have that $\mathbf{E}[M_0^2] \le C$, so $\mathbf{E}[M_k^2] \le C + \sum_{i=1}^k \mathbf{E}[(M_i - M_{i-1})^2]$. Therefore,

$$\sup_k \mathbf{E}[M_k^2] \le C + \sup_k \sum_{i=1}^k \mathbf{E}[(\alpha_i(1 - \nu_i\beta_i)\langle\nabla F(x^i), \xi^i\rangle)^2] \le C + G^2 \sum_{i=1}^{\infty} \mathbf{E}[\alpha_i^2||\xi^i||^2]$$

where the last inquality used Assumption 1.2 and the fact that $0 \leq \nu_i \beta_i \leq 1$ almost surely. Moreover,

$$\sum_{i=1}^{\infty} \mathbf{E}[\alpha_i^2 ||\xi^i||^2] = \sum_{i=1}^{\infty} \mathbf{E}[\mathbf{E}_i[\alpha_i^2||\xi^i||^2]] = \sum_{i=1}^{\infty} \mathbf{E}[\alpha_i^2 \mathbf{E}_i[||\xi^i||^2] \leq C_2 \sum_{i=1}^{\infty} \mathbf{E}[\alpha_i^2]$$

where the first equality is by the law of total expectation and the inequality comes from Assumption 1.3. Because $\sum_{i=1}^{\infty} \alpha_i^2 < \infty$, we finally have

$$\sup_k \mathbf{E}[M_k^2] < \infty$$

so M_k is a convergent martingale. In particular,

$$\sum_{k=0}^{\infty} -\alpha_k(1 - \nu_k \beta_k)\langle \nabla F(x^k), \xi^k \rangle > -\infty \ a.s.$$

Combining this with Lemma 3, we get

$$\sum_{k=0}^{\infty} W_k < \infty \ a.s.$$

We have shown (26) and (27), which concludes the proof. □

Now we prove Theorem 2, where under a stronger noise assumption we show that $\beta_k \to 1$ is admissible as long as it goes to 1 slow enough.

Proof of Theorem 2. Assume the sequences $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\nu_k\}$ satisfy the following:

$$\sum_{k=0}^{\infty} \alpha_k = \infty \tag{33}$$

$$\sum_{k=0}^{\infty} (1 - \nu_k \beta_k)^2 < \infty \tag{34}$$

$$\sum_{k=0}^{\infty} \frac{\alpha_k^2}{1 - \nu_k \beta_k} < \infty \tag{35}$$

$$\lim_{k \to \infty} \beta_k = 1 \tag{36}$$

then sequence $\{x^k\}$ generated by the algorithm (6) satisfies

$$\liminf_{k \to \infty} ||\nabla F(x^k)|| = 0 \quad a.s \tag{37}$$

By the smoothness assumption, we have

$$\begin{aligned} F(x^{k+1}) &\leq F(x^k) + \langle \nabla F(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}||x^{k+1} - x^k||^2 \\ &= F(x^k) + \langle \nabla F(x^k), -\alpha_k b^k \rangle + \frac{L}{2}\alpha_k^2||b^k||^2 \end{aligned}$$

$$
\begin{aligned}
&= F(x^k) + \left\langle \nabla F(x^k),\, -\alpha_k\left(\nabla F(x^k) + b^k - \nabla F(x^k)\right)\right\rangle + \frac{L}{2}\alpha_k^2\|b^k\|^2 \\
&= F(x^k) - \alpha_k\|\nabla F(x^k)\|^2 - \alpha_k\langle \nabla F(x^k),\, b^k - \nabla F(x^k)\rangle + \frac{L}{2}\alpha_k^2\|b^k\|^2 \quad (38)
\end{aligned}
$$

Using the update formula in (6), we have

$$
\begin{aligned}
b^k - \nabla F(x^k) &= (1-\nu_k)g^k + \nu_k d^k - \nabla F(x^k) \\
&= (1-\nu_k\beta_k)g^k + \nu_k\beta_k d^{k-1} - \nabla F(x^k) \\
&= (1-\nu_k\beta_k)\big(g^k - \nabla F(x^k)\big) + \nu_k\beta_k\big(d^{k-1} - \nabla F(x^k)\big) \\
&= (1-\nu_k\beta_k)\xi^k + \nu_k\beta_k\big(d^{k-1} - \nabla F(x^k)\big) \quad (39)
\end{aligned}
$$

Substitution of (39) into (38) yields

$$
\begin{aligned}
F(x^{k+1}) &\leq F(x^k) - \alpha_k\|\nabla F(x^k)\|^2 - \alpha_k\nu_k\beta_k\langle \nabla F(x^k),\, d^{k-1} - \nabla F(x^k)\rangle \\
&\quad -\alpha_k(1-\nu_k\beta_k)\langle \nabla F(x^k),\, \xi^k\rangle + \frac{L}{2}\alpha_k^2\|b^k\|^2 \\
&\leq F(x^k) - \alpha_k\|\nabla F(x^k)\|^2 + \alpha_k\nu_k\beta_k\left(\frac{1}{4}\|F(x^k)\|^2 + \|d^{k-1} - \nabla F(x^k)\|^2\right) \\
&\quad -\alpha_k(1-\nu_k\beta_k)\langle \nabla F(x^k),\, \xi^k\rangle + \frac{L}{2}\alpha_k^2\|b^k\|^2 \\
&\leq F(x^k) - \frac{3\alpha_k}{4}\|\nabla F(x^k)\|^2 + \alpha_k\nu_k\beta_k\|d^{k-1} - \nabla F(x^k)\|^2 \\
&\quad -\alpha_k(1-\nu_k\beta_k)\langle \nabla F(x^k),\, \xi^k\rangle + \frac{L}{2}\alpha_k^2\|b^k\|^2 \quad (40)
\end{aligned}
$$

where in the second inequality we used $\langle a, b\rangle \leq \frac{1}{4}\|a\|^2 + \|b\|^2$ for any two vectors a and b, and in the last inequality we used $0 \leq \nu_k\beta_k \leq 1$. Taking conditional expectation on both sides of the above inequality and using $\mathbf{E}[\xi^k] = 0$, we get

$$
\mathbf{E}_k\big[F(x^{k+1})\big] \leq F(x^k) - \frac{3\alpha_k}{4}\|\nabla F(x^k)\|^2 + \alpha_k\nu_k\beta_k\|d^{k-1} - \nabla F(x^k)\|^2 + \frac{L}{2}\alpha_k^2\mathbf{E}_k\big[\|b^k\|^2\big] \quad (41)
$$

Next we analyze the sequence $\{d^{k-1} - \nabla F(x^k)\}$. From the update formula in (6), we have

$$
\begin{aligned}
d^k - \nabla F(x^{k+1}) &= \nu_k\beta_k d^{k-1} + (1-\nu_k\beta_k)g^k - \nabla F(x^{k+1}) + \nabla F(x^k) - \nabla F(x^k) \\
&= \nu_k\beta_k\big(d^{k-1} - \nabla F(x^k)\big) + (1-\nu_k\beta_k)\big(g^k - \nabla F(x^k)\big) + \big(\nabla F(x^k) - \nabla F(x^{k+1})\big) \\
&= \nu_k\beta_k\big(d^{k-1} - \nabla F(x^k)\big) + (1-\nu_k\beta_k)\xi^k + \big(\nabla F(x^k) - \nabla F(x^{k+1})\big)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\|d^k - \nabla F(x^{k+1})\|^2 &= (\nu_k\beta_k)^2\|d^{k-1} - \nabla F(x^k)\|^2 + \|(1-\nu_k\beta_k)\xi^k + \big(\nabla F(x^k) - \nabla F(x^{k+1})\big)\|^2 \\
&\quad + 2\nu_k\beta_k\left\langle d^{k-1} - \nabla F(x^k),\, (1-\nu_k\beta_k)\xi^k + \big(\nabla F(x^k) - \nabla F(x^{k+1})\big)\right\rangle \\
&\leq (\nu_k\beta_k)^2\|d^{k-1} - \nabla F(x^k)\|^2 + 2(1-\nu_k\beta_k)^2\|\xi^k\|^2 + 2\|\nabla F(x^k) - \nabla F(x^{k+1})\|^2 \\
&\quad + 2\nu_k\beta_k\left\langle d^{k-1} - \nabla F(x^k),\, (1-\nu_k\beta_k)\xi^k\right\rangle \\
&\quad + 2\nu_k\beta_k\left\langle d^{k-1} - \nabla F(x^k),\, \big(\nabla F(x^k) - \nabla F(x^{k+1})\big)\right\rangle
\end{aligned}
$$

21

$$\leq (\nu_k \beta_k)^2 \big\| d^{k-1} - \nabla F(x^k) \big\|^2 + 2(1 - \nu_k \beta_k)^2 \big\| \xi^k \big\|^2 + 2 \big\| \nabla F(x^k) - \nabla F(x^{k+1}) \big\|^2$$

$$+ 2\nu_k \beta_k \left\langle d^{k-1} - \nabla F(x^k), \, (1 - \nu_k \beta_k) \xi^k \right\rangle$$

$$+ \nu_k \beta_k \left(\eta \big\| d^{k-1} - \nabla F(x^k) \big\|^2 + \frac{1}{\eta} \big\| \nabla F(x^k) - \nabla F(x^{k+1}) \big\|^2 \right),$$

where in the first inequality we used $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, and in the second inequality we used $2\langle a, b\rangle \leq \eta \|a\|^2 + \frac{1}{\eta}\|b\|^2$ for any $\eta > 0$. By the smoothness assumption, we have

$$\big\| \nabla F(x^k) - \nabla F(x^{k+1}) \big\|^2 \leq L^2 \big\| x^k - x^{k+1} \big\|^2 = \alpha_k^2 L^2 \big\| b^k \big\|^2$$

which, combining with the previous inequality, leads to

$$\big\| d^k - \nabla F(x^{k+1}) \big\|^2 \leq \left((\nu_k \beta_k)^2 + \nu_k \beta_k \eta \right) \big\| d^{k-1} - \nabla F(x^k) \big\|^2 + 2(1 - \nu_k \beta_k)^2 \big\| \xi^k \big\|^2 + 2\alpha_k^2 L^2 \big\| b^k \big\|^2$$

$$+ 2\nu_k \beta_k \left\langle d^{k-1} - \nabla F(x^k), \, (1 - \nu_k \beta_k) \xi^k \right\rangle + \frac{\nu_k \beta_k}{\eta} \alpha_k^2 L^2 \big\| b^k \big\|^2$$

$$\leq \nu_k \beta_k (\nu_k \beta_k + \eta) \big\| d^{k-1} - \nabla F(x^k) \big\|^2 + 2(1 - \nu_k \beta_k)^2 \big\| \xi^k \big\|^2 + 2\alpha_k^2 L^2 \big\| b^k \big\|^2$$

$$+ 2\nu_k \beta_k \left\langle d^{k-1} - \nabla F(x^k), \, (1 - \nu_k \beta_k) \xi^k \right\rangle + \frac{1}{\eta} \alpha_k^2 L^2 \big\| b^k \big\|^2$$

Choosing $\eta = 1 - \nu_k \beta_k$, we obtain

$$\big\| d^k - \nabla F(x^{k+1}) \big\|^2 \leq \nu_k \beta_k \big\| d^{k-1} - \nabla F(x^k) \big\|^2 + 2(1 - \nu_k \beta_k)^2 \big\| \xi^k \big\|^2 + 2\alpha_k^2 L^2 \big\| b^k \big\|^2$$

$$+ 2\nu_k \beta_k \left\langle d^{k-1} - \nabla F(x^k), \, (1 - \nu_k \beta_k) \xi^k \right\rangle + \frac{\alpha_k^2}{1 - \nu_k \beta_k} L^2 \big\| b^k \big\|^2 \quad (42)$$

Taking expectation conditioned on $\{x^0, g^0, \ldots, x^{k-1}, d^{k-1}, x^k\}$, we have

$$\mathbf{E}_k \big[\big\| d^k - \nabla F(x^{k+1}) \big\|^2 \big] \leq \nu_k \beta_k \big\| d^{k-1} - \nabla F(x^k) \big\|^2 + 2(1 - \nu_k \beta_k)^2 \mathbf{E}_k \big[\big\| \xi^k \big\|^2 \big] + 2\alpha_k^2 L^2 \mathbf{E}_k \big[\big\| b^k \big\|^2 \big]$$

$$+ \frac{\alpha_k^2}{1 - \nu_k \beta_k} L^2 \mathbf{E}_k \big[\big\| b^k \big\|^2 \big] \quad (43)$$

To show that $\|d^{k-1} - \nabla F(x^k)\|^2$ is a convergent martingale, we prove the following lemma, similar to [Ermoliev(1969)].

Lemma 4. *Assume we are given a sequence such that $\mathbf{E}_k[X_{k+1}] \leq X_k + Y_k$, where $0 \leq X_k \leq C$ and $0 \leq Y_k \leq C$ almost surely for some constant C, the random variables Y_k are $\mathcal{F}_k$-measurable, and they satisfy $\sum_{k=0}^\infty Y_k < \infty$ almost surely. Then the sequence X_k converges almost surely.*

Proof. We show that $Z_k = X_k + \sum_{k=0}^\infty Y_k$ is a convergent supermartingale. By Doob decomposition, Z_k is a supermartingale if and only if the sequence

$$A_k = \sum_{i=1}^k \mathbf{E}_{i-1}[Z_i - Z_{i-1}]$$

satisfies $\mathbf{P}(A_{k+1} \leq A_k) = 1$ for all k [Williams(1991), section 12.11]. Here

$$\mathbf{E}_{i-1}[Z_i - Z_{i-1}] = \mathbf{E}_{i-1}[X_k] - X_{k-1} - Y_{k-1}$$

and we assumed this is non-positive. So $A_{k+1} \leq A_k$ almost surely. The upper bound on X_k and the convergence of $\sum_{k=0}^\infty Y_k$ implies that the supermartingale Z is in $\mathcal{L}^1$, so the sequence $\{Z_k\}$ converges almost surely by Doob's forward convergence theorem [Williams(1991), chapter 11]. By the convergence of $\sum Y_k$, this in turn implies that the sequence X_k converges almost surely. $\square$

We can apply the above lemma to show that $\left\|d^{k-1}-\nabla F(x^k)\right\|^2$ is a convergent semimartingale. Because the noise $\|\xi^k\|^2$ is uniformly bounded almost surely and $\|\nabla F(x^k)\| \le G$, $\|d^{k-1}-\nabla F(x^k)\|^2$ is uniformly bounded in k. The uniform bound on the noise $\|\xi^k\|^2$ also implies that $\|b^k\|^2$ is uniformly bounded in k. In the notation of the lemma, we have

$$Y_k = 2(1-\nu_k\beta_k)^2\mathbf{E}_k\big[\|\xi^k\|^2\big] + 2\alpha_k^2 L^2\mathbf{E}_k\big[\|b^k\|^2\big] + \frac{\alpha_k^2}{1-\nu_k\beta_k}L^2\mathbf{E}_k\big[\|b^k\|^2\big]$$

Note that $Y_k \ge 0$. To show convergence of $\sum Y_k$, note that the uniform bounds imply

$$Y_k \le C\left((1-\nu_k\beta_k)^2 + \alpha_k^2 + \frac{\alpha_k^2}{1-\nu_k\beta_k}\right)$$

for suitably large C. Then convergence follows from the conditions on the sequences $(1-\nu_k\beta_k)^2$, α_k^2, and $\alpha_k^2/(1-\nu_k\beta_k)$. This proves that $\|d^{k-1}-\nabla F(x^k)\|^2$ converges almost surely.

Summing up the two inequalities (41) and (43) gives

$$\mathbf{E}_k\left[F(x^{k+1}) + \left\|d^k - \nabla F(x^{k+1})\right\|^2\right]$$

$$\le \quad F(x^k) + (1+\alpha_k)\nu_k\beta_k\left\|d^{k-1}-\nabla F(x^k)\right\|^2 - \frac{3\alpha_k}{4}\left\|\nabla F(x^k)\right\|^2$$

$$+2(1-\nu_k\beta_k)^2\mathbf{E}_k\big[\|\xi^k\|^2\big] + \frac{5}{2}\alpha_k^2 L^2\mathbf{E}_k\big[\|b^k\|^2\big] + \frac{\alpha_k^2}{1-\nu_k\beta_k}L^2\mathbf{E}_k\big[\|b^k\|^2\big]$$

If $(1+\alpha_k)\nu_k\beta_k \le 1$, then

$$\mathbf{E}_k\left[F(x^{k+1}) + \left\|d^k - \nabla F(x^{k+1})\right\|^2\right]$$

$$\le \quad F(x^k) + \left\|d^{k-1}-\nabla F(x^k)\right\|^2 - \frac{3\alpha_k}{4}\left\|\nabla F(x^k)\right\|^2$$

$$+2(1-\nu_k\beta_k)^2\mathbf{E}_k\big[\|\xi^k\|^2\big] + \frac{5}{2}\alpha_k^2 L^2\mathbf{E}_k\big[\|b^k\|^2\big] + \frac{\alpha_k^2}{1-\nu_k\beta_k}L^2\mathbf{E}_k\big[\|b^k\|^2\big]$$

Rearranging terms, we get

$$\frac{3\alpha_k}{4}\left\|\nabla F(x^k)\right\|^2 \quad \le \quad F(x^k) + \left\|d^{k-1}-\nabla F(x^k)\right\|^2 - \mathbf{E}_k\left[F(x^{k+1}) + \left\|d^k - \nabla F(x^{k+1})\right\|^2\right]$$

$$+2(1-\nu_k\beta_k)^2\mathbf{E}_k\big[\|\xi^k\|^2\big] + \frac{5}{2}\alpha_k^2 L^2\mathbf{E}_k\big[\|b^k\|^2\big] + \frac{\alpha_k^2}{1-\nu_k\beta_k}L^2\mathbf{E}_k\big[\|b^k\|^2\big]$$

Since $\alpha_k/(1-\nu_k\beta_k) \to 0$, there exists m such that $(1+\alpha_k)\nu_k\beta_k \le 1$ for all $k \ge m$. Taking full expectation on both sides of the above inequality and summing up for all $k \ge m$, we obtain

$$\frac{3}{4}\sum_{k=m}^{\infty}\mathbf{E}\left[\alpha_k\left\|\nabla F(x^k)\right\|^2\right] \quad \le \quad \mathbf{E}\left[F(x^m) + \left\|d^{m-1}-\nabla F(x^m)\right\|^2\right] - F_*$$

$$+\sum_{k=m}^{\infty}2C_\xi\mathbf{E}\big[\beta_k^2\big] + \sum_{k=m}^{\infty}\frac{5}{2}L^2 C_d\mathbf{E}\big[\alpha_k^2\big] + \sum_{k=m}^{\infty}L^2 C_d\mathbf{E}\left[\frac{\alpha_k^2}{\beta_k}\right]$$

$$\le \quad M + C\left(\sum_{k=m}^{\infty}(1-\nu_k\beta_k)^2 + \sum_{k=m}^{\infty}\alpha_k^2 + \sum_{k=m}^{\infty}\frac{\alpha_k^2}{(1-\nu_k\beta_k)}\right)$$

The right-hand side is bounded by assumption (the index m is finite), so we have

$$\frac{3}{4} \sum_{k=m}^{\infty} \mathbf{E} \left[\alpha_k \|\nabla F(x^k)\|^2 \right] < \infty$$

This in turn implies that the series

$$\frac{3}{4} \sum_{k=m}^{\infty} \alpha_k \|\nabla F(x^k)\|^2 < \infty \qquad a.s.$$

So because $\sum_k \alpha_k = \infty$, there must be a subsequence k_t with $\|\nabla F(x^{k_t})\|^2 \to 0$. This proves (7). $\qquad \square$

C Local Convergence Rate Proofs

In this section we give a proof to Theorem 3 and it's generalized version, which we present below.

We will denote with $\lambda_i(A)$, $\rho(A)$ the i-th eigenvalue and spectral radius of the matrix A respectively.

Let's recall the equations of deterministic QHM algorithm (6) with constant parameters α, β, ν:

$$\begin{aligned} d^k &= (1 - \beta)\nabla F(x^k) + \beta d^{k-1} \\ x^{k+1} &= x^k - \alpha \left[(1 - \nu)\nabla F(x^k) + \nu d^k \right] \end{aligned} \tag{44}$$

In this section we will assume that d^0 is initialized with zero vector.

Taking the gradient of the quadratic function $F(x) = x^T A x + b^T x + c$ and substituting it into (44) yields

$$\begin{aligned} d^k &= (1 - \beta)(Ax^k + b) + \beta d^{k-1} \\ x^{k+1} &= x^k - \alpha(1 - \nu\beta)(Ax^k + b) - \alpha\nu\beta d^{k-1} \end{aligned}$$

Plugging in $Ax_* = -b$ we get

$$\begin{aligned} d^k &= (1 - \beta)A(x^k - x_*) + \beta d^{k-1} \\ x^{k+1} - x_* &= x^k - x_* - \alpha(1 - \nu\beta)A(x^k - x_*) - \alpha\nu\beta d^{k-1} \end{aligned}$$

We can write the above two equations as

$$\begin{bmatrix} d^k \\ x^{k+1} - x_* \end{bmatrix} = \begin{bmatrix} \beta I & (1 - \beta)A \\ -\alpha\nu\beta I & I - \alpha(1 - \nu\beta)A \end{bmatrix} \begin{bmatrix} d^{k-1} \\ x^k - x_* \end{bmatrix} \triangleq T(\theta) \begin{bmatrix} d^{k-1} \\ x^k - x_* \end{bmatrix} = T^k(\theta, A) \begin{bmatrix} d^0 \\ x^0 - x_* \end{bmatrix}$$

where I denotes the $n \times n$ identity matrix and $\theta = \{\alpha, \beta, \nu\}$. It is known that the sequence of $T^k(\theta, A)$ converges to zero if and only if the spectral radius $\rho(T) < 1$. Moreover, Gelfand's Formula states that $\rho(T) = \lim_{k \to \infty} \|T^k\|^{\frac{1}{k}}$, which means that $\exists \{\epsilon_k\}_0^\infty$, $\lim_{k \to \infty} \epsilon_k = 0$ such that

$$\left\| x^{k+1} - x_* \right\| \le \left\| \begin{bmatrix} d^k \\ x^{k+1} - x_* \end{bmatrix} \right\| \le \left\| T^k(\theta, A) \right\| \left\| \begin{bmatrix} d^0 \\ x^0 - x_* \end{bmatrix} \right\| \le (\rho(T) + \epsilon_k)^k \left\| x^0 - x_* \right\|$$

Thus, the behavior of the algorithm is determined by the eigenvalues of $T(\theta)$. To find them, we will use a standard technique of changing basis. Let $A = Q\Lambda Q^T$ be an eigendecomposition of the matrix A. Then, multiplying A with Q and appropriate permutation matrix P^6 we get

$$
P \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} T(\theta) \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix}^T P^T = \begin{bmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T_n \end{bmatrix}
$$

where $T_i \in \mathbb{R}^{2\times 2}$ is defined as

$$
T_i = T_i(\theta, \lambda_i(A)) = \begin{bmatrix} \beta & (1-\beta)\lambda_i(A) \\ -\alpha\nu\beta & 1 - \alpha(1 - \nu\beta)\lambda_i(A) \end{bmatrix}
$$

Thus, to compute eigenvalues of T, it is enough to compute the eigenvalues of all matrices T_i. We use the following Lemma to establish the region when $\rho(T_i) < 1$:

Lemma 5. *Let* $\alpha > 0, \beta \in [0, 1), \nu \in [0, 1], \lambda_i(A) > 0$. *Then*

$$
\rho(T_i(\theta)) < 1 \ \text{if} \ \alpha < \frac{2(1+\beta)}{\lambda_i(A)(1 + \beta(1 - 2\nu))}
$$

Proof. Let's denote with λ eigenvalues of T_i. Let's also define $l \triangleq \lambda_i(A)$. Then, λ satisfies the following equation:

$$
\begin{aligned}
&(\beta - \lambda)\left(1 - \alpha\left(1 - \nu\beta\right)l - \lambda\right) + \alpha\nu\beta\left(1 - \beta\right)l = 0 \Leftrightarrow \\
&\beta - \lambda - \beta\alpha(1 - \nu\beta)l + \lambda\alpha(1 - \nu\beta)l - \lambda\beta + \lambda^2 + \alpha\nu\beta l - \alpha\nu\beta^2 l = 0 \Leftrightarrow \\
&\beta - \lambda - \beta\alpha l + \alpha\beta^2\nu l + \lambda\alpha l - \lambda\alpha\nu\beta l - \lambda\beta + \lambda^2 + \alpha\beta\nu l - \alpha\beta^2\nu l = 0 \Leftrightarrow \\
&\lambda^2 - (1 - \alpha l + \alpha\nu\beta l + \beta)\lambda + \beta(1 - \alpha l + \alpha\nu l) = 0 \\
&D = (1 - \alpha l + \alpha\nu\beta l + \beta)^2 - 4\beta(1 - \alpha l + \alpha\nu l)
\end{aligned}
$$

Let's denote by $S(A) = \{\alpha, \beta, \nu : A \text{ is true}\}$. The final convergence set

$$
S(|\lambda| < 1) = S(|\lambda| < 1 \cap D \geq 0) \cup S(|\lambda| < 1 \cap D < 0)
$$

Let's look at the case when $D \geq 0$. Then $S(|\lambda| < 1 \cap D \geq 0) = S(D \geq 0) \cap S(|\lambda_1| < 1) \cap S(|\lambda_2| < 1)$

$$
\lambda_{1,2} = \frac{1 - \alpha l + \alpha\nu\beta l + \beta \pm \sqrt{D}}{2}
$$

Let's look at $S(|\lambda_1| < 1) = S(\lambda_1 < 1) \cap S(\lambda_1 > -1)$

$$
\begin{aligned}
|\lambda_1| = \frac{\left|1 - \alpha l + \alpha\nu\beta l + \beta + \sqrt{D}\right|}{2} < 1 &\Leftrightarrow \left|1 - \alpha l + \alpha\nu\beta l + \beta + \sqrt{D}\right| < 2 \Leftrightarrow \\
-2 < 1 - \alpha l + \alpha\nu\beta l + \beta + \sqrt{D} < 2 &\Leftrightarrow
\end{aligned}
$$

---

[6]See, e.g. [Recht(2010)] for the exact form of matrix P.

$$-3 + \alpha l - \alpha \nu \beta l - \beta < \sqrt{D} < 1 + \alpha l - \alpha \nu \beta l - \beta$$

Let's solve the second inequality: $S(\lambda_1 < 1)$. Since we are only interested in the case when $D \geq 0$ we get

$$\sqrt{D} < 1 + \alpha l - \alpha \nu \beta l - \beta \Leftrightarrow (1 - \alpha l + \alpha \nu \beta l + \beta)^2 - 4\beta(1 - \alpha l + \alpha \nu l) < (1 + \alpha l - \alpha \nu \beta l - \beta)^2 \Leftrightarrow$$
$$0 < 4\beta(1 - \alpha l + \alpha \nu l) + 4(\alpha l - \alpha \nu \beta l - \beta) \Leftrightarrow 0 < 4\alpha l(1 - \beta)$$

which is always satisfied.

Let's solve the first inequality:

$$S(\lambda_1 > -1) = S(-3 + \alpha l - \alpha \nu \beta l - \beta \leq 0) \cup S(-3 + \alpha l - \alpha \nu \beta l - \beta > 0 \cap -3 + \alpha l - \alpha \nu \beta l - \beta < \sqrt{D})$$

We can rewrite the first term as

$$S(-3 + \alpha l - \alpha \nu \beta l - \beta \leq 0) = S\left(\alpha \leq \frac{3 + \beta}{l(1 - \nu \beta)}\right)$$

Let's compute the second term

$$(\alpha l - \alpha \nu \beta l - \beta - 3)^2 < (\alpha l - \alpha \nu \beta l - \beta - 1)^2 - 4\beta(1 - \alpha l + \alpha \nu l) \Leftrightarrow$$
$$0 < 4(\alpha l - \alpha \nu \beta l - \beta) - 8 - 4\beta(1 - \alpha l + \alpha \nu l) \Leftrightarrow$$
$$0 < -8\beta - 8\alpha \nu \beta l + 4\alpha l(1 + \beta) - 8 \Leftrightarrow$$
$$2(1 + \beta) < \alpha l(1 + \beta - 2\beta \nu) \Leftrightarrow \alpha > \frac{2(1 + \beta)}{l(1 + \beta(1 - 2\nu))}$$

The last inequality is true since $1 + \beta(1 - 2\nu) > 0$. Thus

$$S(-3 + \alpha l - \alpha \nu \beta l - \beta > 0 \cap -3 + \alpha l - \alpha \nu \beta l - \beta < \sqrt{D}) =$$
$$S\left(\alpha > \frac{3 + \beta}{l(1 - \nu \beta)}\right) \cap S\left(\alpha > \frac{2(1 + \beta)}{l(1 + \beta(1 - 2\nu))}\right) =$$
$$S\left(\alpha > \frac{3 + \beta}{l(1 - \nu \beta)}\right)$$

Since $3 + \beta > 2(1 + \beta)$ and $l(1 - \nu \beta) \leq l(1 + \beta(1 - 2\nu))$.

Therefore we have that $\lambda_1 > -1$ always holds and thus $|\lambda_1| < 1$ always holds.

Now we compute $S(|\lambda_2| < 1) = S(\lambda_2 < 1) \cap S(\lambda_2 > -1)$. The first term is

$$\lambda_2 = \frac{1 - \alpha l + \alpha \nu \beta l + \beta - \sqrt{D}}{2} < 1 \Leftrightarrow 1 - \alpha l + \alpha \nu \beta l + \beta - \sqrt{D} < 2 \Leftrightarrow -1 - \alpha l + \alpha \nu \beta l + \beta < \sqrt{D}$$

Which is always satisfied since

$$-1 - \alpha l + \alpha \nu \beta l + \beta = \beta - 1 + l\alpha(\nu \beta - 1) < 0$$

Let's compute the second term:

$$\lambda_2 = \frac{1 - \alpha l + \alpha \nu \beta l + \beta - \sqrt{D}}{2} > -1 \Leftrightarrow$$

26

$$1 - \alpha l + \alpha \nu \beta l + \beta - \sqrt{D} > -2 \Leftrightarrow \sqrt{D} < 3 - \alpha l + \alpha \nu \beta l + \beta \Leftrightarrow$$

$$(1 - \alpha l + \alpha \nu \beta l + \beta)^2 - 4\beta(1 - \alpha l + \alpha \nu l) < (3 - \alpha l + \alpha \nu \beta l + \beta)^2 \Leftrightarrow$$

$$-8 + 4(\alpha l - \alpha \nu \beta l - \beta) - 4\beta(1 - \alpha l + \alpha \nu l) < 0 \Leftrightarrow \alpha < \frac{2(1 + \beta)}{l(1 + \beta(1 - 2\nu))}$$

Thus we get

$$S(|\lambda_2| < 1) = S\left(\alpha < \frac{2(1 + \beta)}{l(1 + \beta(1 - 2\nu))}\right)$$

and therefore

$$S(|\lambda| < 1 \cap D > 0) = S(D > 0) \cap S\left(\alpha < \frac{2(1 + \beta)}{l(1 + \beta(1 - 2\nu))}\right)$$

Now let's move to the second case and compute $S(|\lambda| < 1 \cap D < 0)$. If $D < 0$ we have that $1 - \alpha l + \alpha \nu l > 0$ and then

$$|\lambda_{1,2}| = \frac{|1 - \alpha l + \alpha \nu \beta l + \beta \pm i\sqrt{-D}|}{2} = 0.5\sqrt{(1 - \alpha l + \alpha \nu \beta l + \beta)^2 - D} =$$

$$\sqrt{\beta(1 - \alpha l + \alpha \nu l)} < 1 \Leftrightarrow \alpha > \frac{\beta - 1}{l\beta(1 - \nu)}$$

which is always true, so

$$S(|\lambda| < 1 \cap D < 0) = S(D < 0)$$

Finally, let's find a simplified form of $S(D \geq 0)$ and $S(D < 0)$.

$$D = (1 - \alpha l + \alpha \nu \beta l + \beta)^2 - 4\beta(1 - \alpha l + \alpha \nu l) = (1 + \beta - \alpha l(1 - \nu\beta))^2 - 4\beta + 4\alpha l\beta - 4\alpha l\beta\nu$$

$$= 1 + 2\beta + \beta^2 - 2\alpha l(1 + \beta)(1 - \nu\beta) + \alpha^2 l^2(1 - \nu\beta)^2 - 4\beta + 4\alpha l\beta - 4\alpha l\beta\nu$$

$$= \alpha^2 l^2(1 - \nu\beta)^2 - 2\alpha l - 2\alpha l\beta + 2\alpha l\nu\beta + 2\alpha l\nu\beta^2 + 4\alpha l\beta - 4\alpha l\beta\nu + 1 - 2\beta + \beta^2$$

$$= \alpha^2 l^2(1 - \nu\beta)^2 - 2\alpha l + 2\alpha l\beta - 2\alpha l\nu\beta + 2\alpha l\nu\beta^2 + (1 - \beta)^2$$

$$= \alpha^2 l^2(1 - \nu\beta)^2 - 2\alpha l(1 - \beta + \nu\beta - \nu\beta^2) + (1 - \beta)^2$$

$$= \alpha^2 l^2(1 - \nu\beta)^2 - 2\alpha l(1 - \beta)(1 + \nu\beta) + (1 - \beta)^2$$

Let's denote the discriminant of that equation (divided by 4) with respect to αl as D_1:

$$D_1 = (1 - \beta)^2(1 + \nu\beta)^2 - (1 - \nu\beta)^2(1 - \beta)^2 = 4\nu\beta(1 - \beta)^2 \geq 0$$

$$\alpha l_{1,2} = \frac{(1 - \beta)(1 + \nu\beta) \pm 2(1 - \beta)\sqrt{\nu\beta}}{(1 - \nu\beta)^2} = \frac{(1 - \beta)(1 + \nu\beta \pm 2\sqrt{\nu\beta})}{(1 - \nu\beta)^2} = \frac{(1 - \beta)(1 \pm 2\sqrt{\nu\beta})^2}{(1 - \nu\beta)^2}$$

Therefore

$$S(D \geq 0) = S\left(\alpha \geq \frac{(1 - \beta)(1 + \sqrt{\nu\beta})^2}{l(1 - \nu\beta)^2} \cup \alpha \leq \frac{(1 - \beta)(1 - \sqrt{\nu\beta})^2}{l(1 - \nu\beta)^2}\right)$$

$$S(D < 0) = S\left(\alpha \in \left[\frac{(1 - \beta)(1 - \sqrt{\nu\beta})^2}{l(1 - \nu\beta)^2}, \frac{(1 - \beta)(1 + \sqrt{\nu\beta})^2}{l(1 - \nu\beta)^2}\right]\right)$$

(45)

Now, notice that

$$\frac{(1-\beta)(1+\sqrt{\nu\beta})^2}{l(1-\nu\beta)^2} = \frac{(1+\sqrt{\nu\beta})^2}{\frac{l(1-\nu\beta)^2}{1-\beta}} < \frac{2(1+\beta)}{l(1+\beta(1-2\nu))}$$

since $(1+\sqrt{\nu\beta})^2 < 2(1+\beta)$ (left side is less than 2 and right side is greater than 2) and also

$$\frac{l(1-\nu\beta)^2}{1-\beta} \geq l(1+\beta(1-2\nu)) \Leftrightarrow 1 - 2\nu\beta + \nu^2\beta^2 \geq 1 - \beta + \beta - \beta^2 - 2\nu\beta + 2\nu\beta^2 \Leftrightarrow \beta^2(1-\nu)^2 \geq 0$$

Thus overall we get that

$$
\begin{aligned}
S(|\lambda| < 1 \cap D \geq 0) &= S(D \geq 0) \cap S\left(\alpha < \frac{2(1+\beta)}{l(1+\beta(1-2\nu))}\right) = \\
&= S\left(\alpha < \frac{2(1+\beta)}{l(1+\beta(1-2\nu))}\right) \setminus S(D < 0) \Rightarrow \\
S(|\lambda| < 1) &= S(|\lambda| < 1 \cap D \geq 0) \cup S(|\lambda| < 1 \cap D < 0) \\
&= S\left(\alpha < \frac{2(1+\beta)}{l(1+\beta(1-2\nu))}\right) \setminus S(D < 0) \cup S(D < 0) = \\
&= S\left(\alpha < \frac{2(1+\beta)}{l(1+\beta(1-2\nu))}\right)
\end{aligned}
$$

$\square$

Now, let's establish a precise equation for the spectral radius $\rho(T_i)$.

Lemma 6. *Let $\alpha > 0, \beta \in [0,1), \nu \in [0,1], \lambda_i(A) > 0$. Let's define $l \triangleq \lambda_i(A)$ and*

$$
\begin{aligned}
C_1 &= 1 - \alpha l + \alpha l \nu \beta + \beta \\
C_2 &= \beta(1 - \alpha l + \alpha l \nu)
\end{aligned}
$$

Then

$$
r(\theta, l) = \rho(T_i(\theta)) = \begin{cases}
0.5\left(\sqrt{C_1^2 - 4C_2} + C_1\right) & \text{if } C_1 \geq 0, C_1^2 - 4C_2 \geq 0 \\
0.5\left(\sqrt{C_1^2 - 4C_2} - C_1\right) & \text{if } C_1 < 0, C_1^2 - 4C_2 \geq 0 \\
\sqrt{C_2} & \text{if } C_1^2 - 4C_2 < 0
\end{cases}
$$

In addition, $r(\theta, l)$ is non-increasing as a function of l for $0 < l < \frac{1-\beta}{\alpha(1-\sqrt{\nu\beta})^2}$ and is non-decreasing for $l > \frac{1-\beta}{\alpha(1-\sqrt{\nu\beta})^2}$.

Proof. Following derivations from the proof of Lemmas 5 we get

$$
r(\theta, l) = \begin{cases}
\max\left\{0.5\left|C_1 + \sqrt{C_1^2 - 4C_2}\right|, 0.5\left|C_1 - \sqrt{C_1^2 - 4C_2}\right|\right\} & \text{if } C_1^2 - 4C_2 \geq 0 \\
\sqrt{C_2} & \text{if } C_1^2 - 4C_2 < 0
\end{cases}
$$

28

Considering 4 cases for different signs of C_1 and C_2, the first statement of the Lemma immediately follows. To prove the second statement, let's define the following 3 points:

$$p_1 = \frac{1-\beta}{\alpha(1+\sqrt{\nu\beta})^2}, \quad p_2 = \frac{1-\beta}{\alpha(1-\sqrt{\nu\beta})^2}, \quad p_3 = \frac{1+\beta}{\alpha(1-\nu\beta)}$$

From equation (45) and definition of C_1 we get that

$$C_1 \geq 0 \Leftrightarrow l \leq p_3$$
$$C_1^2 - 4C_2 \geq 0 \Leftrightarrow l \leq p_1 \text{ or } l \geq p_2$$

and it is easy to check that if $\beta \geq \nu \Rightarrow p_1 \leq p_2 \leq p_3$ and if $\beta < \nu \Rightarrow p_1 \leq p_3 \leq p_2$. Moreover, both $C_1(l)$ and $C_2(l)$ are non-increasing function of l, and $C_1^2(l) - 4C_2(l)$ is non-increasing when $l \leq p_2$ and non-decreasing when $l \geq p_1$.

Let's first prove the second statement of the Lemma for the case when $\beta < \nu$. In that case, when $l < p_1$ the function is non-increasing, since both $C_1(l)$ and $C_1^2(l) - 4C_2(l)$ are non-increasing. When $p_1 \leq l \leq p_2$, the function is non-increasing, because $C_2(l)$ is non-increasing. Finally, when $l > p_2$, the function is non-decreasing, because both $C_1^2(l) - 4C_2(l)$ and $-C_1(l)$ are non-decreasing.

When $\beta \geq \nu$, the same reasoning applies, but we additionally need to prove that the function is non-decreasing when $p_2 \leq l \leq p_3$. In that case $r(\theta, l) = 0.5(\sqrt{C_1^2(l) - 4C_2(l)} + C_1(l))$. Taking the derivative of r with respect to l we get

$$\frac{\partial r}{\partial l} = \frac{2\alpha\beta(1-\nu) - \alpha(1-\nu\beta)\left(\sqrt{C_1^2(l) - 4C_2(l)} + C_1(l)\right)}{2\sqrt{C_1^2(l) - 4C_2(l)}}$$

Let's show that this derivative is always non-negative when $l \geq p_2$

$$\frac{2\alpha\beta(1-\nu) - \alpha(1-\nu\beta)\left(\sqrt{C_1^2(l) - 4C_2(l)} + C_1(l)\right)}{2\sqrt{C_1^2(l) - 4C_2(l)}} \geq 0 \Leftrightarrow$$

$$2\beta(1-\nu) - (1-\nu\beta)\left(\sqrt{C_1^2(l) - 4C_2(l)} + C_1(l)\right) \geq 0 \Leftrightarrow$$

$$4\beta(1-\nu)^2 - 4\beta(1-\nu)(1-\nu\beta)C_1(l) + (1-\nu\beta)C_1^2(l) \geq (C_1^2(l) - 4C_2(l))(1-\nu\beta)^2 \Leftrightarrow$$

$$\beta(1-\nu)^2 - \beta(1-\nu)(1-\nu\beta)C_1(l) + C_2(l)(1-\nu\beta)^2 \geq 0 \Leftrightarrow$$

$$\frac{-\beta^2(1-\nu^2) + (1-\nu)(1-\nu\beta)(1+\beta) - \beta(1-\nu\beta)^2}{\alpha(1-\nu)(1-\nu\beta)^2(1-\beta)} \leq l \Leftrightarrow$$

$$-\frac{(1-\beta)^2\nu\beta}{\alpha(1-\nu)(1-\nu\beta)^2(1-\beta)} \leq l$$

which is always true since left side is less than zero.

$\square$

The last thing that we need in order to prove Theorem 3 is given by the following Lemma:

Lemma 7. *Let* $\mu \leq \min_i \lambda_i(A)$ *and* $L \geq \max_i \lambda_i(A)$. *Then*

$$\rho(T(\theta)) \leq R(\theta, \mu, L) = \max\left(r(\theta, \mu), r(\theta, L)\right)$$

In addition, the minimal spectral radius with respect to θ depends on μ and L only through κ, i.e. $\min_\theta R(\theta, \mu, L) \triangleq R^*(\kappa)$

Proof. To prove this first statement of the Lemma, let's notice that by definition

$$\rho(T(\theta)) = \max_i \rho(T_i(\theta))$$

But Lemma 6 states that $\rho(T_i(\theta))$ is first non-increasing and then non-decreasing with respect to the eigenvalues of A. Thus, the maximum can only be achieved on the boundaries, which are precisely equal to or smaller than $r(\theta, \mu)$ and $r(\theta, L)$.

Let's prove the second statement of the Lemma by contradiction. Let's assume that the optimal rate does in fact depend on μ and L not only through κ. That means that $\exists \mu_1, L_1, \mu_2, L_2$, such that $L_1/\mu_1 = L_2/\mu_2$, but $\min_\theta R(\theta, \mu_1, L_1) \neq \min_\theta R(\theta, \mu_2, L_2)$. Let's consider the optimal rates if the function f is divided by μ_1 for the first case and by μ_2 for the second. In that case, $\min_\theta R(\theta, 1, L_1/\mu_1) = \min_\theta R(\theta, 1, L_2/\mu_2)$. But on the other hand, they can't be equal, since we have that $\min_\theta R(\theta, 1, L_1/\mu_1) = \min_\theta R(\theta, \mu_1, L_1)$ and $\min_\theta R(\theta, 1, L_2/\mu_2) = \min_\theta R(\theta, \mu_2, L_2)$, because multiplying learning rate by μ_1 for the first case and by μ_2 for the second yields exactly the same sequence of iterates and thus the optimal rate can't change. □

Now we are ready to prove Theorem 3. We restate it below for convenience

Theorem 3. *Let's denote $\theta = \{\alpha, \beta, \nu\}$. For any function $F(x) = x^T A x + b^T x + c$ that satisfies $\mu \leq \lambda_i(A) \leq L$ for all $i = 1, \ldots, n$ and any x^0, the deterministic QHM algorithm $z^{k+1} = Tz^k$ satisfies*

$$\left\| x^k - x_* \right\| \leq (R(\theta, \mu, L) + \epsilon_k)^k \left\| x^0 - x_* \right\|,$$

where $x_* = \arg\min_x F(x)$, $\lim_{k\to\infty} \epsilon_k = 0$ and $R(\theta, \mu, L) = \rho(T)$, which can be characterized as

$$R(\theta, \mu, L) = \max \left\{ r(\theta, \mu), r(\theta, L) \right\}, \quad where$$

$$r(\theta, \lambda) = \begin{cases} 0.5 \left(\sqrt{C_1(\lambda)^2 - 4C_2(\lambda)} + C_1(\lambda) \right) & \text{if } C_1(\lambda) \geq 0, C_1(\lambda)^2 - 4C_2(\lambda) \geq 0, \\ 0.5 \left(\sqrt{C_1(\lambda)^2 - 4C_2(\lambda)} - C_1(\lambda) \right) & \text{if } C_1(\lambda) < 0, C_1(\lambda)^2 - 4C_2(\lambda) \geq 0, \\ \sqrt{C_2(\lambda)} & \text{if } C_1(\lambda)^2 - 4C_2(\lambda) < 0, \end{cases}$$

$$C_1(\lambda, \theta) = 1 - \alpha\lambda + \alpha\lambda\nu\beta + \beta,$$
$$C_2(\lambda, \theta) = \beta(1 - \alpha\lambda + \alpha\lambda\nu)$$

To ensure $R(\theta, \mu, L) < 1$, the parameters α, β, ν must satisfy the following constraints:

$$0 < \alpha < \frac{2(1+\beta)}{L(1 + \beta(1 - 2\nu))}, \quad 0 \leq \beta < 1, \quad 0 \leq \nu \leq 1$$

In addition, the optimal rate depends only on κ: $\min_\theta R(\theta, \mu, L)$ is a function of only κ.

Proof. Lemma 6 and Lemma 7 immediately give the first statement of the Theorem. One can also get the bound on the function values by using definition of the Lipschitz continuous gradient:

$$F(x^k) - F(x_*) \leq \nabla F(x_*)^T (x^k - x_*) + \frac{L}{2} \left\| x^k - x_* \right\|^2 = \frac{L}{2} \left\| x^k - x_* \right\|^2$$

Finally, to get the stability region, we apply Lemma 5 and notice that $\lambda_i(A) \leq L \; \forall i$. □

To generalize this result, let's define the following class of functions

Definition 1. $\mathcal{F}_{\mu,L}^1$ *is the class of all functions* $F : \mathbb{R}^n \to \mathbb{R}$ *that are continuously differentiable, strongly convex with parameter* μ *and have Lipschitz continuous gradient with parameter* L. *We will denote the condition number of* F *as* $\kappa = L/\mu$.

Then, Theorem 3 can be generalized to any function $F \in \mathcal{F}_{\mu,L}^1$ in the following way:

Theorem 6. *Let's denote* $\theta = \{\alpha, \beta, \nu\}$. *For any function* $F \in \mathcal{F}_{\mu,L}^1$ *that is additionally twice differentiable at the point* $x_* = \arg\min_x F(x)$, *deterministic QHM algorithm locally converges to* x_* *with linear rate, from any initialization* x^0 *sufficiently close to* x_*.
Precisely, for any $\epsilon \in [0, 1 - R(\theta, \mu, L)) \; \exists \; \delta > 0$ *and* $c \geq 0$, *such that* $\forall k \geq 0$ *the following holds*

$$\left\| x^k - x_* \right\| \leq c \left(R(\theta, \mu, L) + \epsilon \right)^k$$

$$F(x^k) - F(x_*) \leq \frac{c^2 L}{2} \left(R(\theta, \mu, L) + \epsilon \right)^{2k}$$

$$R(\theta, \mu, L) = \max \left\{ r(\theta, \mu), r(\theta, L) \right\}$$

$$r(\theta, \lambda) = \begin{cases} 0.5 \left(\sqrt{C_1(\lambda)^2 - 4C_2(\lambda)} + C_1(\lambda) \right) & \text{if } C_1(\lambda) \geq 0, C_1(\lambda)^2 - 4C_2(\lambda) \geq 0 \\ 0.5 \left(\sqrt{C_1(\lambda)^2 - 4C_2(\lambda)} - C_1(\lambda) \right) & \text{if } C_1(\lambda) < 0, C_1(\lambda)^2 - 4C_2(\lambda) \geq 0 \\ \sqrt{C_2(\lambda)} & \text{if } C_1(\lambda)^2 - 4C_2(\lambda) < 0 \end{cases}$$

$$C_1(\lambda, \theta) = 1 - \alpha\lambda + \alpha\lambda\nu\beta + \beta$$

$$C_2(\lambda, \theta) = \beta(1 - \alpha\lambda + \alpha\lambda\nu)$$

if $\left\| x^0 - x_* \right\| \leq \delta$ *and* α, β, ν *satisfy the following constraints:*

$$0 < \alpha < \frac{2(1 + \beta)}{L(1 + \beta(1 - 2\nu))}, 0 \leq \beta < 1, 0 \leq \nu \leq 1$$

In addition, the optimal rate depends on μ *and* L *only through* κ, *i.e.* $\min_\theta R(\theta, \mu, L) \triangleq R^*(\kappa)$.

Proof. To prove this result we apply Lyapunov's method (see e.g. Chapter 2, Theorem 1 of [Polyak(1987)]) to the QHM equations. The proof is then identical to the proof of Theorem 3, with matrix A replaced by $\nabla^2 F(x_*)$. $\qquad \square$

D Numerical Evaluation of the Convergence Rate

In this section we provide details on the numerical evaluation of the local convergence rate of QHM. We need to numerically estimate the following function

$$R^*(\nu, \kappa) = \min_{\alpha, \beta} \max \{ r(\alpha, \beta, \nu, \mu), r(\alpha, \beta, \nu, L) \}$$

From Lemma 6 (Appendix C) we know that $r(\alpha, \beta, \nu, l)$ is a non-increasing function of l until some point and non-decreasing after. Also note that in fact dependence of r on α is the same as on l, since they only appear in formulas as a product αl. Thus, it is easy to see that for optimal α we will have

$$r(\alpha, \beta, \nu, \mu) = r(\alpha, \beta, \nu, L) \tag{46}$$
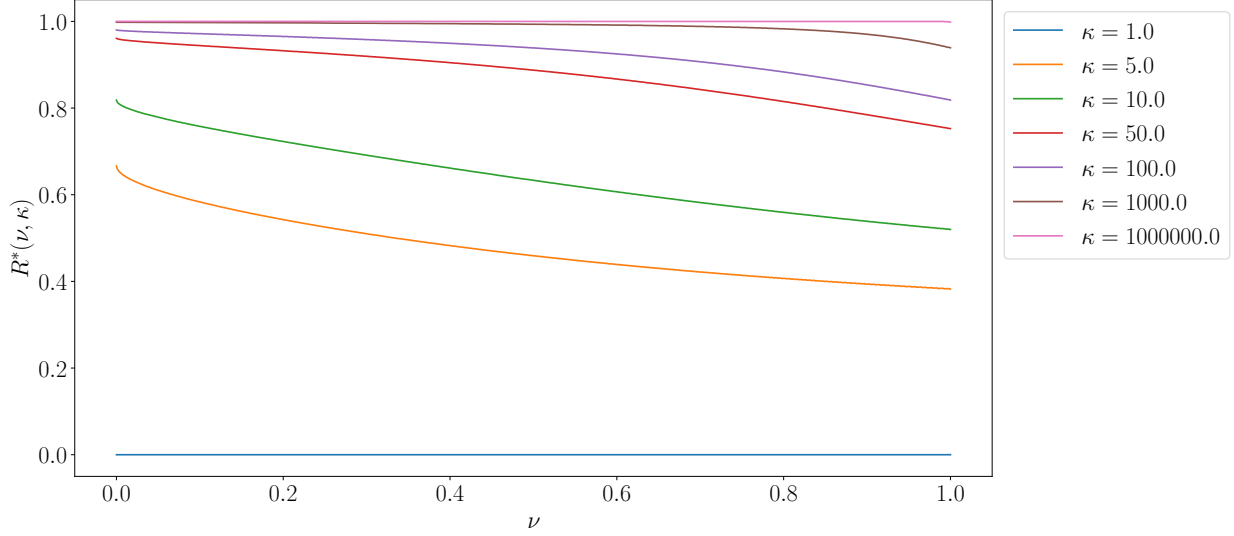
Figure 5. This Figure shows the dependence of optimal (across α, β) local convergence rate on ν for QHM algorithm across different values of condition number κ. Note that $R^*(\nu, \kappa)$ always shows monotonic non-decreasing dependence on ν.

because otherwise α could be changed to decrease the value of the maximum.

Thus, to find optimal α for fixed β, ν, we can solve equation (46) for α using binary search (with precision set to 10^{-8}). To find optimal β or ν we just use grid search (with grid size equal to 10^3) on $[0, 1 - 10^{-5}]$ for β and $[0, 1]$ for ν.

To numerically verify that the dependence of the optimal rate on ν is monotonic, we run this procedure for 10^3 values of κ which are sampled (on a uniform grids) in the following way: 100 values on $[1, 10]$, 100 values on $[10, 100]$, 100 values on $[100, 1000]$, 150 values on $[10^3, 10^4]$, 150 values on $[10^4, 10^5]$, 200 values on $[10^5, 10^6]$, 200 values on $[10^6, 10^7]$. All experiments were run in parallel using GNU Parallel Command-Line Tool [Tange et al.(2011)].

Since rate estimation is non-exact, it happens sometimes that very close points ν show non-monotonic rate dependence, but it is always the case that the rate is approximately non-increasing in ν. Precisely, we verify that the following condition holds for all estimated values of κ:

$$\overline{R}^*(\nu_{i+10}, \kappa) - \overline{R}^*(\nu_i, \kappa) < 10^{-3} \; \forall i = 1, \dots, 990$$

where $\overline{R}^*$ is estimated rate and ν_i is i-th sample of ν. Figure 5 shows the dependence of $R^*(\nu, \kappa)$ on ν for different values of κ.

E   Stationary Distribution Proofs

In this section we present proofs of Theorems 4, 5. We will restate the combined statement of both theorems below for convenience.

Theorems 4, 5. *Suppose $F(x) = \frac{1}{2}x^T A x$, where A is symmetric positive definite matrix. The stochastic gradients satisfy $g^k = \nabla F(x^k) + \xi$, where ξ is a random vector independent of x^k with zero mean $\mathbf{E}[\xi] = 0$ and covariance matrix $\mathbf{E}[\xi\xi^T] = \Sigma_\xi$. Also suppose the parameteres α, β, ν*

satisfy (13), then QHM algorithm (6), equivalently (10) in this case, converges to stationary distribution satisfying

$$A\Sigma_x + \Sigma_x A = \alpha A\Sigma_\xi + O(\alpha^2)$$

$$\mathbf{tr}(A\Sigma_x) = \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\left(1 + \frac{2\nu\beta}{1-\beta}\left[\frac{2\nu\beta}{1+\beta} - 1\right]\right)\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3)$$

Consequently, when $\nu = 0$ (SGD), Σ_x satisfies

$$\mathbf{tr}(A\Sigma_x) = \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3)$$

When $\nu = 1$ (SHB), Σ_x satisfies

$$\mathbf{tr}(A\Sigma_x) = \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\frac{1-\beta}{1+\beta}\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3)$$

When $\nu = \beta$ (NAG), Σ_x satisfies

$$\mathbf{tr}(A\Sigma_x) = \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\left(1 - \frac{2\beta^2(1+2\beta)}{1+\beta}\right)\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3)$$

Proof. We consider the behavior of QHM with constant α, β, and ν, described in (47).

$$\begin{aligned} d^k &= (1-\beta)g^k + \beta d^{k-1} \\ x^{k+1} &= x^k - \alpha\left[(1-\nu)g^k + \nu d^k\right] \end{aligned} \tag{47}$$

Under assumptions of Theorems 4, 5 we have that stochastic gradient g^k is generated as

$$g^k = \nabla F(x^k) + \xi^k = Ax^k + \xi^k \tag{48}$$

where the noise ξ^k is independent of x^k, has zero mean and constant covariance matrix. More explicitly, for all $k \geq 0$,

$$\mathbf{E}\left[\xi^k\right] = 0, \qquad \mathbf{E}\left[\xi^k(\xi^k)^T\right] = \Sigma_\xi$$

where Σ_ξ is a constant covariance matrix. Substituting the expression of g^k in (48) into (47) yields

$$\begin{aligned} d^k &= (1-\beta)g^k + \beta d^{k-1} = (1-\beta)Ax^k + (1-\beta)\xi^k + \beta d^{k-1}, \\ x^{k+1} &= x^k - \alpha\nu d^k - \alpha(1-\nu)g^k = x^k - \alpha(1-\nu\beta)Ax^k - \alpha\nu\beta d^{k-1} - \alpha(1-\nu\beta)\xi^k \end{aligned}$$

We can write the above two equations as

$$\begin{bmatrix} d^k \\ x^{k+1} \end{bmatrix} = \begin{bmatrix} \beta I & (1-\beta)A \\ -\alpha\nu\beta I & I - \alpha(1-\nu\beta)A \end{bmatrix} \begin{bmatrix} d^{k-1} \\ x^k \end{bmatrix} + \begin{bmatrix} (1-\beta)I \\ -\alpha(1-\nu\beta)I \end{bmatrix} \xi^k \tag{49}$$

where I denotes the $n \times n$ identity matrix.

Let $L > 0$ be the largest eigenvalue of A. From Theorem 3 we know that under the conditions (13) the dynamical system (49) is stable, i.e., the spectral radius of the matrix

$$\begin{bmatrix} \beta I & (1-\beta)A \\ -\alpha\nu\beta I & I - \alpha(1-\nu\beta)A \end{bmatrix}$$

33

is smaller than one.

To simplify notation, we rewrite Equation (49) as

$$z^{k+1} = z^k - Bz^k + C\xi^k \tag{50}$$

where

$$z^k = \begin{bmatrix} d^{k-1} \\ x^k \end{bmatrix}, \qquad B = \begin{bmatrix} (1-\beta)I & -(1-\beta)A \\ \alpha\nu\beta I & \alpha(1-\nu\beta)A \end{bmatrix}, \qquad C = \begin{bmatrix} (1-\beta)I \\ -\alpha(1-\nu\beta)I \end{bmatrix}$$

As $k \to \infty$, the effect of the initial point z^0 dies out and the covariance matrix of the state z^k becomes constant. Let

$$\Sigma_z = \begin{bmatrix} \Sigma_d & \Sigma_{dx} \\ \Sigma_{xd} & \Sigma_x \end{bmatrix} \triangleq \lim_{k\to\infty} \mathbf{E}\big[z^k(z^k)^T\big] = \lim_{k\to\infty} \begin{bmatrix} \mathbf{E}\big[d^{k-1}(d^{k-1})^T\big] & \mathbf{E}\big[d^{k-1}(x^k)^T\big] \\ \mathbf{E}\big[x^k(d^{k-1})^T\big] & \mathbf{E}\big[x^k(x^k)^T\big] \end{bmatrix}$$

Then using the linear dynamics (50) and the assumption that $\{\xi^k\}$ is i.i.d. and has zero mean, we obtain

$$B\Sigma_z + \Sigma_z B^T - B\Sigma_z B^T = C\Sigma_\xi C^T$$

Following the partition of Σ_z, we partition the above matrix equation into 2 by 2 blocks and obtain

$$(1,1): \quad (1-\beta^2)\Sigma_d - \beta(1-\beta)(A\Sigma_{xd} + \Sigma_{dx}A) - (1-\beta)^2 A\Sigma_x A = (1-\beta)^2\Sigma_\xi \tag{51}$$

$$(1,2): \quad \alpha\nu\beta^2\Sigma_d + (1-\beta)\Sigma_{dx} - (1-\beta)A\Sigma_x + \alpha\nu\beta(1-\beta)A\Sigma_{xd} + \alpha\beta(1-\nu\beta)\Sigma_{dx}A$$
$$\qquad +\alpha(1-\nu\beta)(1-\beta)A\Sigma_x A = -\alpha(1-\nu\beta)(1-\beta)\Sigma_\xi \tag{52}$$

$$(2,1): \quad \alpha\nu\beta^2\Sigma_d + (1-\beta)\Sigma_{xd} - (1-\beta)\Sigma_x A + \alpha\beta(1-\nu\beta)A\Sigma_{xd} + \alpha\nu\beta(1-\beta)\Sigma_{dx}A$$
$$\qquad +\alpha(1-\nu\beta)(1-\beta)A\Sigma_x A = -\alpha(1-\nu\beta)(1-\beta)\Sigma_\xi \tag{53}$$

$$(2,2): \quad -(\alpha\nu\beta)^2\Sigma_d + \alpha\nu\beta(\Sigma_{xd} + \Sigma_{dx}) + \alpha(1-\nu\beta)(A\Sigma_x + \Sigma_x A)$$
$$\qquad -\alpha^2\nu\beta(1-\nu\beta)(A\Sigma_{xd} + \Sigma_{dx}A) - \alpha^2(1-\nu\beta)^2 A\Sigma_x A = \alpha^2(1-\nu\beta)^2\Sigma_\xi \tag{54}$$

Or, letting V be the column block matrix with entries $[\Sigma_d, \Sigma_{xd}, \Sigma_{dx}, A\Sigma_x, \Sigma_x A, A\Sigma_{xd}, \Sigma_{dx}A, A\Sigma_x A]$, and defining symbolically U to be the block matrix with coefficients:

$$\begin{bmatrix} (1-\beta^2) & 0 & 0 & 0 & 0 & -\beta(1-\beta) & -\beta(1-\beta) & -(1-\beta)^2 \\ \alpha\nu\beta^2 & 0 & (1-\beta) & -(1-\beta) & 0 & \alpha\nu\beta(1-\beta) & \alpha\beta(1-\nu\beta) & \alpha(1-\nu\beta)(1-\beta) \\ \alpha\nu\beta^2 & (1-\beta) & 0 & 0 & -(1-\beta) & \alpha\beta(1-\nu\beta) & \alpha\nu\beta(1-\beta) & \alpha(1-\nu\beta)(1-\beta) \\ -(\alpha\nu\beta)^2 & \alpha\nu\beta & \alpha\nu\beta & \alpha(1-\nu\beta) & \alpha(1-\nu\beta) & -\alpha^2\nu\beta(1-\nu\beta) & -\alpha^2\nu\beta(1-\nu\beta) & -\alpha^2(1-\nu\beta)^2, \end{bmatrix} \tag{55}$$

(each block is an $n \times n$ identity matrix), we have

$$UV = \begin{bmatrix} (1-\beta)^2 \\ -\alpha(1-\nu\beta)(1-\beta) \\ -\alpha(1-\nu\beta)(1-\beta) \\ \alpha^2(1-\nu\beta)^2 \end{bmatrix} \Sigma_\xi \tag{56}$$

Next we use combinations of the above equations to obtain simplified relations: First, we can do

$$\frac{\alpha^2(1-\nu\beta)^2}{(1-\beta)^2}(1,1) + \frac{\alpha(1-\nu\beta)}{1-\beta}[(1,2) + (2,1)] + (2,2)$$

to get

$$\frac{\alpha(1 + \beta - 2\nu\beta)}{1 - \beta}\Sigma_d + \Sigma_{xd} + \Sigma_{dx} = 0$$

We take the following asymptotic expansion of Σ_z:

$$\begin{bmatrix} \Sigma_d & \Sigma_{dx} \\ \Sigma_{xd} & \Sigma_x \end{bmatrix} = \begin{bmatrix} \Sigma_d^{(0)} + \alpha\Sigma_d^{(1)} + \alpha^2\Sigma_d^{(2)}/2 & \alpha\Sigma_{dx}^{(1)} + \alpha^2\Sigma_{dx}^{(2)}/2 \\ \alpha\Sigma_{xd}^{(1)} + \alpha^2\Sigma_{xd}^{(2)}/2 & \alpha\Sigma_x^{(1)} + \alpha^2\Sigma_x^{(2)}/2 \end{bmatrix} \tag{57}$$

Here, we explicitly write the zero'th order of $(\Sigma_{dx}, \Sigma_{xd}, \Sigma_x)$ to be zero. This can be easily proved from (51)-(54).

The zero'th order term of (51) gives

$$(1 - \beta^2)\Sigma_d^{(0)} = (1 - \beta)^2\Sigma_\xi \tag{58}$$

The first order term of (52) (and (53)) gives

$$\nu\beta^2\Sigma_d^{(0)} + (1 - \beta)(\Sigma_{dx}^{(1)} - A\Sigma_x^{(1)}) = -(1 - \nu\beta)(1 - \beta)\Sigma_\xi \tag{59}$$

$$\nu\beta^2\Sigma_d^{(0)} + (1 - \beta)(\Sigma_{xd}^{(1)} - \Sigma_x^{(1)}A) = -(1 - \nu\beta)(1 - \beta)\Sigma_\xi \tag{60}$$

The second order term of (54) gives

$$\nu\beta(\Sigma_{xd}^{(1)} + \Sigma_{dx}^{(1)}) + (1 - \nu\beta)(A\Sigma_x^{(1)} + \Sigma_x^{(1)}A) = \nu^2\beta^2\Sigma_d^{(0)} + (1 - \nu\beta)^2\Sigma_\xi \tag{61}$$

From (58) we solve

$$\Sigma_d^{(0)} = \frac{1 - \beta}{1 + \beta}\Sigma_\xi,$$

and from (58), (59) and (60), we solve

$$\Sigma_{xd}^{(1)} + \Sigma_{dx}^{(1)} = A\Sigma_x^{(1)} + \Sigma_x^{(1)}A - \frac{2(1 + \beta - \nu\beta)}{1 + \beta}\Sigma_\xi \tag{62}$$

After plugging them into (61), we obtain

$$A\Sigma_x^{(1)} + \Sigma_x^{(1)}A = \Sigma_\xi \tag{63}$$

thus

$$A\Sigma_x + \Sigma_x A = \alpha\Sigma_\xi + O(\alpha^2)$$

which concludes the proof of Theorem 4.

Let's now extend this result to the second-order in α. The first order term of (51) gives

$$(1 + \beta)\Sigma_d^{(1)} = \beta(A\Sigma_{xd}^{(1)} + \Sigma_{dx}^{(1)}A) + (1 - \beta)A\Sigma_x^{(1)}A \tag{64}$$

The second order term of (52) (and (53)) gives

$$\nu\beta^2\Sigma_d^{(1)} + \frac{1 - \beta}{2}(\Sigma_{dx}^{(2)} - A\Sigma_x^{(2)}) = -\beta(1 - \nu\beta)\Sigma_{dx}^{(1)}A - \nu\beta(1 - \beta)A\Sigma_{xd}^{(1)} - (1 - \nu\beta)(1 - \beta)A\Sigma_x^{(1)}A \tag{65}$$

$$\nu\beta^2\Sigma_d^{(1)} + \frac{1 - \beta}{2}(\Sigma_{xd}^{(2)} - \Sigma_x^{(2)}A) = -\beta(1 - \nu\beta)A\Sigma_{xd}^{(1)} - \nu\beta(1 - \beta)\Sigma_{dx}^{(1)}A - (1 - \nu\beta)(1 - \beta)A\Sigma_x^{(1)}A \tag{66}$$

The third order term of (54) gives

$$-\nu^2\beta^2\Sigma_d^{(1)}+\frac{\nu\beta}{2}(\Sigma_{xd}^{(2)}+\Sigma_{dx}^{(2)})+\frac{1-\nu\beta}{2}(A\Sigma_x^{(2)}+\Sigma_x^{(2)}A)=\nu\beta(1-\nu\beta)(\Sigma_{dx}^{(1)}A+A\Sigma_{xd}^{(1)})+(1-\nu\beta)^2A\Sigma_x^{(1)}A \tag{67}$$

Plugging (64) into (65) and (66), we obtain

$$(\Sigma_{xd}^{(2)}+\Sigma_{dx}^{(2)})-(A\Sigma_x^{(2)}+\Sigma_x^{(2)}A)=-\frac{2\beta(1+\nu+\beta-\nu\beta)}{1-\beta^2}(\Sigma_{dx}^{(1)}A+A\Sigma_{xd}^{(1)})-\frac{4(1+\beta-\nu\beta)}{1+\beta}A\Sigma_x^{(1)}A \tag{68}$$

Plugging (64) into (67), we obtain

$$\nu\beta(\Sigma_{xd}^{(2)}+\Sigma_{dx}^{(2)})+(1-\nu\beta)(A\Sigma_x^{(2)}+\Sigma_x^{(2)}A)=\frac{2\nu\beta(1+\beta-\nu\beta)}{1+\beta}(\Sigma_{dx}^{(1)}A+A\Sigma_{xd}^{(1)})+2(1-2\nu\beta+\frac{2\nu^2\beta^2}{1+\beta})A\Sigma_x^{(1)}A \tag{69}$$

Combining (68) and (69), we obtain

$$A\Sigma_x^{(2)}+\Sigma_x^{(2)}A=\frac{2\nu\beta}{1-\beta}(\Sigma_{dx}^{(1)}A+A\Sigma_{xd}^{(1)})+2A\Sigma_x^{(1)}A \tag{70}$$

Let's get an expression for $\mathbf{tr}(A\Sigma_\xi)$. From (70) we get (by taking trace and dividing by 2)

$$\mathbf{tr}(A\Sigma_x^{(2)})=\frac{2\nu\beta}{1-\beta}\mathbf{tr}(\Sigma_{dx}^{(1)}A)+\mathbf{tr}(A\Sigma_x^{(1)}A) \tag{71}$$

From (62) we get (by multiplying by A, taking trace and dividing by 2)

$$\mathbf{tr}(\Sigma_{dx}^{(1)}A)=\mathbf{tr}(A\Sigma_x^{(1)}A)-\frac{(1+\beta-\nu\beta)}{1+\beta}\mathbf{tr}(A\Sigma_\xi) \tag{72}$$

Plugging (72) into (71) we get

$$\begin{aligned}\mathbf{tr}(A\Sigma_x^{(2)})&=\frac{2\nu\beta}{1-\beta}\left(\mathbf{tr}(A\Sigma_x^{(1)}A)-\frac{(1+\beta-\nu\beta)}{1+\beta}\mathbf{tr}(A\Sigma_\xi)\right)+\mathbf{tr}(A\Sigma_x^{(1)}A)\\&=\left(\frac{2\nu\beta}{1-\beta}+1\right)\mathbf{tr}(A\Sigma_x^{(1)}A)-\frac{2\nu\beta}{1-\beta}\frac{(1+\beta-\nu\beta)}{1+\beta}\mathbf{tr}(A\Sigma_\xi)\end{aligned} \tag{73}$$

From (63) we get (by multiplying by A and taking trace)

$$\mathbf{tr}(A\Sigma_x^{(1)}A)=\frac{1}{2}\mathbf{tr}(A\Sigma_\xi) \tag{74}$$

and also by taking trace

$$\mathbf{tr}(A\Sigma_x^{(1)})=\frac{1}{2}\mathbf{tr}(\Sigma_\xi) \tag{75}$$

Finally we get

$$\begin{aligned}\mathbf{tr}(A\Sigma_x)&=\alpha\mathbf{tr}(A\Sigma_x^{(1)})+\frac{\alpha^2}{2}\mathbf{tr}(A\Sigma_x^{(2)})+O(\alpha^3)\\&=\frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi)+\frac{\alpha^2}{2}\left[\left(\frac{2\nu\beta}{1-\beta}+1\right)\mathbf{tr}(A\Sigma_x^{(1)}A)-\frac{2\nu\beta}{1-\beta}\frac{(1+\beta-\nu\beta)}{1+\beta}\mathbf{tr}(A\Sigma_\xi)\right]+O(\alpha^3)\end{aligned}$$

$$= \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{2}\left[\left(\frac{2\nu\beta}{1-\beta}+1\right)\frac{1}{2}\mathbf{tr}(A\Sigma_\xi) - \frac{2\nu\beta}{1-\beta}\frac{(1+\beta-\nu\beta)}{1+\beta}\mathbf{tr}(A\Sigma_\xi)\right] + O(\alpha^3)$$

$$= \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\left[\left(\frac{2\nu\beta}{1-\beta}+1\right) - \frac{4\nu\beta}{1-\beta}\frac{(1+\beta-\nu\beta)}{1+\beta}\right]\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3)$$

$$= \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\mathbf{tr}(A\Sigma_\xi) + \frac{\alpha^2}{4}\frac{2\nu\beta}{1-\beta}\left[1 - \frac{2(1+\beta-\nu\beta)}{1+\beta}\right]\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3)$$

$$= \frac{\alpha}{2}\mathbf{tr}(\Sigma_\xi) + \frac{\alpha^2}{4}\mathbf{tr}(A\Sigma_\xi) + \frac{\alpha^2}{4}\frac{2\nu\beta}{1-\beta}\left[\frac{2\nu\beta}{1+\beta} - 1\right]\mathbf{tr}(A\Sigma_\xi) + O(\alpha^3)$$

The special cases of SGD, SHB and NAG can be straightforwardly obtained by substituting corresponding value of ν into the general formula.

$\square$

F   Evaluation of Stationary Distribution Size

In this section we describe experimental details for evaluation of stationary distribution size on different machine learning problems (Section 5). The first problem we consider is a simple 2-dimensional quadratic function, where we add additive zero-mean Gaussian noise independent of the point x, so that all assumptions of Theorem 5 are fully satisfied. For this function we have

$$\mu = 0.1, L = 10.0, \Sigma_\xi = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}$$

We run the QHM algorithm for 1000 iterations starting at the optimal value and plot final loss as average across all 1000 iterations. We evaluate QHM for the following sweeps of hyperparameters: 30 values of α on a uniform grid on $[0.01, 1.5]$, 30 values of β on a uniform grid on $[0, 0.999]$, 30 values of ν on a uniform grid on $[0, 1]$. For each combination of hyperparameters we verify that $\alpha \to 1, \beta \to 1$ indeed decreases the average loss. However, for smaller values of ν the effect of β is smaller, as expected. The dependence on ν can be described by a quadratic function with minimum at some $\nu < 1$. Note that from formula (15) the dependence on ν is indeed quadratic with optimal ν given by

$$\nu_*(\beta) = \begin{cases} \frac{1+\beta}{4\beta} & , \frac{1}{3} \le \beta < 1 \\ 1 & , 0 \le \beta < \frac{1}{3} \end{cases} \tag{76}$$

From this equation the optimal $\nu_*(\beta) \ge 0.5$ and $\nu_*(\beta) \to 0.5$ as $\beta \to 1$. In the experiments we see the same qualitative behavior, but the optimal value of ν is much closer to 1 than predicted by equation (76).

The second problem we consider is logistic regression on MNIST dataset. We run QHM for 50 epochs with batch size of 128 and weight decay (applied both to weights and biases) of 10^{-4} (thus, $\mu \approx 10^{-4}$). The final loss is averaged across last 1000 batches. We evaluate algorithm for 50 values of $\alpha \in [0.01, 30]$ (log-uniform grid), 20 values of $\beta \in [0, 0.999]$ (uniform grid), 20 values of $\nu \in [0, 1]$ (uniform grid).

The final problem we consider is ResNet-18 on CIFAR-10 dataset. We run QHM with batch size of 256 and weight decay of 10^{-4} (applied only to weights). We run algorithm for 80 epochs with constant parameters and average final loss across last 100 batches. We evaluate $\alpha \in \{0.01, 0.05, 0.1, 0.5, 1.0, 2.0, 3.0, 5.0,$ $\beta \in \{0.0, 0.01, 0.2, 0.5, 0.7, 0.9, 0.99, 0.999\}$. In this experiment we always set $\nu = 1$.
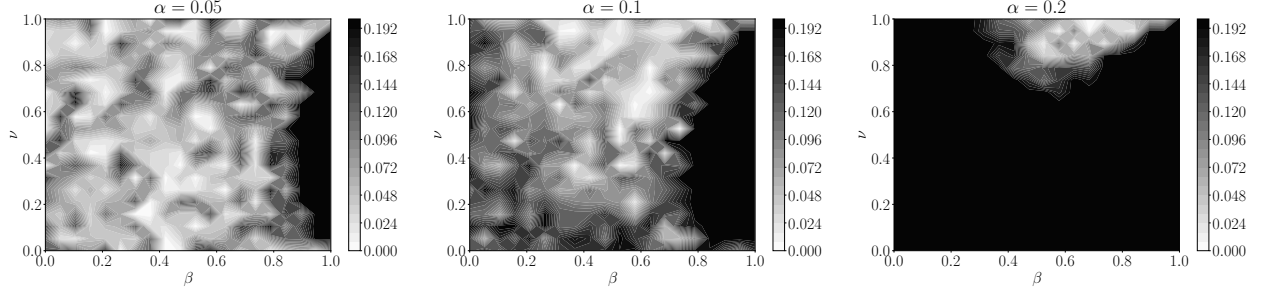
Figure 6. Approximation error of equation (15). Relative error is shown with color and we threshold it at 0.2.

G Approximation error of Theorem 5

In this section we run a set of experiments to check for which values of parameters the equation (15) is not accurate. In fact, we can immediately see that the approximation error grows unboundedly as $\beta \to 1$ if $\nu \notin \{0, \beta, 1\}$, because the right-hand-side of equation (15) converges to $-\infty$, while the left-hand-side is bounded from below.

Since we are interested in the approximation error from the higher-order terms, we run experiments on a 2-dimensional quadratic problem where all assumptions are satisfied. We follow the same experimental settings as in the appendix F. We test a uniform grid of 20 β and 20 ν values on $[0, 1]$ for $\alpha \in \{0.05, 0.1, 0.2\}$. Note, that we can compute the right-hand-side of equation (15) exactly, but need to estimate the left-hand-side. For that we run QHM for 10000 iterations and compute an empirical covariance of the iterates. Figure 6 shows the results of this experiment. We plot a relative error with color and threshold it at 0.2 (i.e. we consider the formula to be inaccurate if the relative difference between right-hand-side and left-hand-side is bigger than 20%). We can see that indeed when α is moderately big, the formula becomes imprecise for many different values of ν and β. However, when α is small, the formula is only imprecise for a very large values of β and it becomes more inaccurate when ν is far from 0 or 1.