ESAIM: Control, Optimisation and Calculus of Variations Will be set by the publisher
URL: `http://www.emath.fr/cocv/`

ALMOST SURE CONVERGENCE OF STOCHASTIC HAMILTONIAN DESCENT METHODS *

Tony Stillfjord[1] and Måns Williamson[2]

Abstract. Gradient normalization and soft clipping are two popular techniques for tackling instability issues and improving convergence of stochastic gradient descent (SGD) with momentum. In this article, we study these types of methods through the lens of dissipative Hamiltonian systems. Gradient normalization and certain types of soft clipping algorithms can be seen as (stochastic) implicit-explicit Euler discretizations of dissipative Hamiltonian systems, where the kinetic energy function determines the type of clipping that is applied. We make use of dynamical systems theory to show in a unified way that all of these schemes converge to stationary points of the objective function, almost surely, in several different settings: a) for L-smooth objective functions, when the variance of the stochastic gradients is possibly infinite, b) under the (L_0, L_1)-smoothness assumption, for heavy-tailed noise with bounded variance, and c) for (L_0, L_1)-smooth functions in the empirical risk minimization setting, when the variance is possibly infinite but the expectation is finite.

2020 Mathematics Subject Classification. 65K10, 46N10, 90C15, 37N40.

The date will be set by the publisher.

1. Introduction

In this article we consider the unconstrained optimization problem

$$\min_{q\in\mathbb{R}^d} F(q), \tag{1}$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is an objective function. A common case in mathematical statistics and machine learning is the empirical risk minimization setting, where F is a weighted sum of loss functions:

$$F(q) = \frac{1}{N}\sum_{i=1}^{N} f_i(q), \tag{2}$$

with $f_i(q) = \ell(h(x_i, q), y_i)$. Here, $\{(x_i, y_i)\}_{i=1}^N$ is an underlying data set of feature-label pairs in the feature-label space $\mathcal{X} \times \mathcal{Y}$, $h(\cdot, q)$ is a model with model parameters q such as a neural network or a regression function, and

Keywords and phrases: Stochastic optimization, Convergence analysis, Clipping, Hamiltonian dynamics

* *The authors were partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The main part of the work was completed while the second author was at Lund University.*

[1] Centre for Mathematical Sciences, Lund University, Box 118, 221 00 Lund, Sweden; e-mail: `tony.stillfjord@math.lth.se`

[2] Kravi Analytics, Scheelevägen 15M, 223 63 Lund, Sweden; e-mail: `mans.williamson@kravianalytics.com`

ℓ is a loss function. A common approach within the machine learning community for solving problems of the type given by (1) is to employ *stochastic gradient descent* (SGD) [47]. A solution q_* to (1) is approximated iteratively by q_k, $k = 1, 2, \ldots$, via

$$q_{k+1} = q_k - \alpha_k \nabla f(q_k, \xi_k), \tag{3}$$

where α_k is the learning rate, $\nabla f(q_k, \xi_k)$ is a stochastic approximation of $\nabla F(q_k)$, and ξ_k is a random variable that accounts for the stochasticity. A common choice is to take a random subset $B_{\xi_k} \subset \{1, \ldots, N\}$ of the indices of the objective function defined by (2) and choose

$$\nabla f(q, \xi_k) = \frac{1}{|B_{\xi_k}|} \sum_{i \in B_{\xi_k}} \nabla f_i(q), \tag{4}$$

where $|B_{\xi_k}|$ denotes the cardinality of B_{ξ_k}. This is attractive when N is very large and $|B_\xi| \ll N$, as it is less computationally expensive than gradient descent. The method also tends to escape local saddle points [16] – an appealing property as many machine learning problems are non-convex. Among the variations of SGD is the popular *SGD with momentum.* Its deterministic counterpart was popularized in the seminal work of Polyak [46]. A common form of this algorithm is expressed as an update in two stages

$$\begin{aligned} p_{k+1} &= \beta_k p_k - \alpha_k \nabla f(q_k, \xi_k), \\ q_{k+1} &= q_k + \alpha_k p_{k+1}, \end{aligned} \tag{5}$$

where $p_0 = 0$ and $\beta_k > 0$ is a momentum parameter. The usage of the momentum update makes the algorithm less sensitive to noise. Indeed, by an iterative argument we obtain that $p_{k+1} = -\sum_{i=0}^{k} \left(\prod_{j=i+1}^{k} \beta_j \right) \alpha_i \nabla f(q_i, \xi_i)$. That is, p_{k+1} is an average of the previous gradients, where β_k determines how much we value information from the preceding stages.

Notwithstanding the benefits of stochastic gradient algorithms, they frequently suffer from instability problems such as exploding gradients [9,42] and sensitivity to the choice of learning rate [41]. A way to mitigate these issues is to employ gradient clipping [24,42] or gradient normalization. Gradient normalization was introduced in [45] in the deterministic setting and a stochastic version appears already in [2]. A normalized version of the algorithm determined by (3) is given by

$$q_{k+1} = q_k - \alpha_k \frac{\nabla f(q_k, \xi_k)}{\|\nabla f(q_k, \xi_k)\|}.$$

In practice a small number $\epsilon > 0$ is added in the denominator to ensure that the update does not become infinitely large.

Gradient clipping was first introduced in [39]. In so-called *hard clipping*, the gradient is simply rescaled if it is larger than some predetermined threshold. *Soft clipping*, on the other hand, makes use of a differentiable function for rescaling the gradient [66]. It was recently shown that hard clipping algorithms suffer from an *unavoidable bias term* [29]; a term in the convergence bound that does not decrease as the number of iterations increases. This is one reason why soft clipping is preferable.

1.1. Gradient normalization, momentum and Hamiltonian systems

In this article, we study gradient normalization and soft clipping of stochastic momentum algorithms from the perspective of *Hamiltonian systems.* As a first step, we note that if we take $\beta_k = 1 - \gamma\alpha_k$ with $\gamma > 0$, we can view the scheme given by (5) as an approximate implicit-explicit Euler discretization of the equation system

$$\begin{aligned} \dot{p} &= -\nabla F(q) - \gamma p, \\ \dot{q} &= p. \end{aligned} \tag{6}$$

The system (6) is *nearly Hamiltonian* [22]; taking

$$H(p,q) = F(q) + \varphi(p), \tag{7}$$

with $\varphi(p) = \frac{1}{2}\|p\|^2$, we can write it on the form

$$\begin{aligned} \dot{p} &= -\nabla_q H(p,q) - \nabla_{\dot{q}} \mathcal{R}(\dot{q}), \\ \dot{q} &= \nabla_p H(p,q). \end{aligned} \tag{8}$$

Here, ∇_p, ∇_q denote the gradients with respect to p and q respectively and $\mathcal{R}(\dot{q}) = \gamma\frac{\|\dot{q}\|^2}{2}$ is a *Rayleigh dissipation function* that accounts for energy dissipation (viscous friction) of the system. Note that this choice of $\mathcal{R}$ yields $\nabla_{\dot{q}}\mathcal{R}(\dot{q}) = \gamma\nabla_p H(p,q)$, which will always be the case in this paper. Thus, for a Hamiltonian of the form (7), (8) reads

$$\begin{aligned} \dot{p} &= -\nabla F(q) - \gamma\nabla\varphi(p), \\ \dot{q} &= \nabla\varphi(p). \end{aligned} \tag{9}$$

We notice that any *fixed point* of this system is a stationary point of F, since $(\dot{q}, \dot{p}) = 0$ implies that $\nabla F(q) = 0$. The dissipation term is often included as an extra term in the Euler-Lagrange equations

$$\nabla_{\dot{q}} L(q,\dot{q}) - \frac{\mathrm{d}}{\mathrm{d}t} L(q,\dot{q}) = \nabla_{\dot{q}}\mathcal{R}(\dot{q}),$$

where $L(q,\dot{q}) = \varphi^*(\dot{q}) - F(q)$ is the *Lagrangian*, and φ^* is the convex conjugate of φ, compare Proposition 51.2 and Example 51.3 in [64]. The physical interpretation is that q is the position of a particle in a potential field $F(q)$ with kinetic energy given by $\varphi^*(\dot{q})$ (in the case when $\varphi(p) = \frac{\|p\|^2}{2}$ we have $\varphi^*(\dot{q}) = \frac{\|\dot{q}\|^2}{2}$). In many scenarios, such as in this case, it happens that the friction term is proportional to the velocity [23]. A ball rolling on a rough incline [11,61]) or on a tilted plane coated with a viscous fluid [10] could for instance be modelled in this fashion, giving weight to the analogy of the *heavy ball* [46]. See also [24], for a further discussion on this.

In this paper, we consider generalizations of the algorithm defined by (5) to equations of the type (8), where φ is a convex, coercive and L-smooth function. The family of schemes we consider are given by

$$\begin{aligned} p_{k+1} &= p_k - \alpha_k\nabla f(q_k,\xi_k) - \alpha_k\gamma\nabla\varphi(p_k), \\ q_{k+1} &= q_k + \alpha_k\nabla\varphi(p_{k+1}), \end{aligned} \tag{10}$$

where $p_0 = 0$, q_0 is arbitrary, and $\{\xi_k\}_{k\geq 0}$ is a sequence of independent, identically distributed random variables. We show that these schemes converge almost surely to the set of stationary points of F. If we take $\varphi(x) = \frac{\|x\|^2}{2}$ in (10), we get SGD with momentum (5). Taking $\varphi(x) = \sqrt{\|x\|^2 + \epsilon}$, $\epsilon > 0$, gives us a gradient normalization scheme, where both the gradient and the momentum variables are rescaled:

$$\begin{aligned} p_{k+1} &= p_k - \alpha_k\nabla f(q_k,\xi_k) - \alpha_k\gamma\frac{p_k}{\sqrt{\|p_k\|^2 + \epsilon}}, \\ q_{k+1} &= q_k + \alpha_k\frac{p_{k+1}}{\sqrt{\|p_{k+1}\|^2 + \epsilon}}. \end{aligned} \tag{11}$$

Other conceivable choices are

i) *Relativistic kinetic energy:* $\varphi(x) = c\sqrt{\|x\|^2 + (mc)^2}$. [19]
ii) *Non-relativistic kinetic energy:* $\varphi(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$, where A is a positive definite, symmetric matrix, $b \in \mathbb{R}^d$ and $c \in \mathbb{R}$. [23]

iii) *Gradient rescaling:* $\varphi(x) = c\sqrt{\|x\|^2 + \epsilon}$, for $c, \epsilon > 0$.
iv) *Soft clipping:* $\varphi(x) = \sqrt{1 + \|x\|^2}$.
v) *The symmetric LogSumExp-function:* $\varphi(x) = \ln\left(\sum_{i=1}^{d} e^{x_i} + e^{-x_i}\right)$, which can be seen as an approximation of the ℓ^∞-norm [52].
vi) *Half-squared ℓ^p-norm:* $\varphi(x) = \frac{1}{2}\|x\|_p^2$, for $p \in [2, \infty)$.

Examples i), iii) and iv) are analytically similar, but give rise to different behaviours in the algorithm given by (10). We refer the reader to [6,44], for verifying that the functions above satisfy the assumptions in Section 3.2.

1.2. Contributions

Making use of Hamiltonian dynamics, we consider a large class of stochastic optimization algorithms (10) for large-scale optimization problems, for which we perform a rigorous convergence analysis. Our assumptions on the dissipation term φ are fairly permissive, and thus the class of algorithms covers both interesting cases like normalized SGD with momentum and various soft-clipping methods with momentum, as well as novel methods. Our analysis shows that the iterates generated by any method in this class are finite almost surely, and that they converge almost surely to the set of stationary points of the objective function F. This means that the methods "always" work in practice, in contrast to what can be guaranteed by analyses that show convergence in expectation. These results are valid in many applications, due to fairly weak assumptions on the optimization problem. The exact assumptions are listed in Section 3 but essentially consist of either

- L-smooth objective functions and stochastic gradients with possibly infinite variance, or
- (L_0, L_1)-smooth objective functions and heavy-tailed stochastic gradients with bounded variance, or
- (L_0, L_1)-smooth objective functions arising in the empirical risk minimization setting and stochastic gradients with possibly infinite variance but bounded expectation.

In particular, we do not assume convexity of the objective function F in any of the cases.

1.3. Outline

In Section 2, we briefly discuss some results that are related to the analysis in this paper. The main results and analysis is presented in Section 3, with conclusions in Section 4. The details of the analysis can be found in Appendix A. This depends on some auxiliary results listed in Appendices B and C.

2. Related works

In this section, we first consider other formulations of SGD with momentum and how the formulation in this paper relates to them. In the second subsection we summarize work in optimization and statistics which make use of Hamiltonian dynamics. Next, we discuss the approach we use for showing almost sure convergence of the methods. Finally, we discuss the central (L_0, L_1)-smoothness condition on the objective function, which is defined later in Assumption 4.

2.1. Momentum algorithms

The implementations of SGD with momentum in the libraries TensorFlow [1] and PyTorch [43] given by

$$p_{k+1} = \beta_k p_k - \alpha_k \nabla f(q_k, \xi_k)$$
$$q_{k+1} = q_k + p_{k+1}$$

are equivalent to (5) after the transformation $\alpha_k \to \sqrt{\alpha_k}$, $\beta_k \to \beta_k\sqrt{\frac{\alpha_{k+1}}{\alpha_k}}$. The method (5) also resembles the (hard-clipped) scheme proposed in [38]:

$$\begin{aligned} p_{k+1} &= \mathrm{clip}_r\left((1-\beta_k)p_k - \beta_k \nabla f(q_k, \xi_k)\right), \\ q_{k+1} &= q_k + \alpha_k p_{k+1}, \end{aligned}$$

where clip_r is a projection operator that projects the argument onto a ball of radius r at the origin. Further, (5) is reminiscent of Stochastic Primal Averaging (SPA) [14]:

$$\begin{aligned} p_{k+1} &= p_k - \eta_k \nabla f(q_k, \xi_k), \\ q_{k+1} &= (1 - c_{k+1})q_k + c_{k+1}p_{k+1}. \end{aligned}$$

In Theorem 1 in [14] it is shown that this is equivalent to the SGD with momentum version

$$\begin{aligned} p_{k+1} &= \beta_k p_k + \nabla f(q_k, \xi_k), \\ q_{k+1} &= q_k - \alpha_k p_{k+1}, \end{aligned}$$

if one takes $\eta_{k+1} = \frac{\eta_k - \alpha_k}{\beta_{k+1}}$ and $c_{k+1} = \frac{\alpha_k}{\eta_k}$. The SPA algorithm can be seen as a randomized implicit-explicit Euler discretization of the equation system

$$\begin{aligned} \dot{p} &= -\nabla F(q), \\ \dot{q} &= p - q, \end{aligned}$$

which after a change of variable is equivalent to (6) for $\gamma = 1$. Under the rather strong assumption that the noise is almost surely bounded (which does not hold for, e.g., Gaussian noise), so-called mixed-clipped SGD with momentum was studied in [66]:

$$\begin{aligned} p_{k+1} &= \beta p_k - (1-\beta)\nabla f(q_k, \xi_k), \\ q_{k+1} &= q_k - \left[\nu \min\left(\eta, \frac{\gamma}{\|p_{k+1}\|}\right) p_{k+1} + (1-\nu)\min\left(\eta, \frac{\gamma}{\|\nabla f(q_k, \xi_k)\|}\right)\nabla f(q_k, \xi_k)\right], \end{aligned}$$

Here, $0 \le \nu \le 1$ is an interpolation parameter.

A drawback with the previously mentioned analyses is that the convergence results are obtained in expectation, which means that there is no guarantee that a single path will converge.

2.2. **Hamiltonian dynamics**

Hamiltonian dynamics, in its energy conserving form, has been well-explored in the Markov chain Monte Carlo field, compare [33]. In [35], various kinetic energy functions φ are considered for equation (8) without the dissipation term $\nabla_{\dot{q}}\mathcal{R}(\dot{q})$.

The algorithm (10) was studied in the context of stochastic differential equations and Langevin dynamics in [54], where the noise is assumed to be Gaussian. In general, however, this is a restrictive assumption in the stochastic optimization setting.

The specific update (11) bears resemblance to deterministic time integration- and optimization schemes studied in [19], that arise as discretizations of the system

$$\begin{aligned} \dot{p} &= -\nabla_q H(p, q) - \gamma p, \\ \dot{q} &= \nabla_p H(p, q), \end{aligned}$$

where the dissipation term γp emanates from *Bateman's Lagrangian* $L(q, \dot{q}) = e^{\gamma t}(\varphi^*(\dot{q}) - F(q))$, see [5]. A similar point of view is also taken in [18], but where so-called Bregman dynamics is employed. In the (deterministic) optimization setting this was studied in [37], where strictly convex kinetic energy functions φ are considered. A stochastic gradient version is analysed in [27] for strongly convex objective functions F.

However, the stochastic optimization algorithm has not been studied for non-convex problems, and an analysis for merely convex (and not strictly convex) kinetic energy functions is lacking.

2.3. Almost sure convergence

The analysis in this paper is based on the *ODE method*, emanating from [36]. The particular proof strategy is due to [30], and is based on linear interpolation of the sequence of iterates. The technique was extended to piecewise constant interpolations in [31]. The approach relies on the assumption that the iterates generated by the algorithm are finite almost surely; an assumption that has to be verified independently.

A similar analysis of the SGD with momentum was performed in [21]. It was extended in [4], to a class of schemes that encompasses (5). The analytical approach is slightly different and does not cover the normalization- and clipping algorithms that we analyze in this article.

We also note that one can employ an analysis similar to that in e.g. [12], along with martingale results like that in [48] to obtain almost sure convergence of a subsequence of the iterates. This is for instance the case in [51] where almost sure convergence guarantees of the type $\min_{0 \le k \le K} \|\nabla F(q_k)\| \to 0$ almost surely for SGD and SGD with momentum are established. These types of results are weaker than those obtained in this paper, since they cannot guarantee that the whole sequence of iterates $\{q_k\}_{k \ge 0}$ converges to a stationary point.

2.4. (L_0, L_1)-smoothness

The (L_0, L_1)-smoothness assumption (defined later in Assumption 4) was introduced in [67] as a more appropriate measure of smoothness than the classic L-smoothness for certain machine learning problems. It is shown in [67] that the iteration complexity of clipped SGD is bounded, under the assumption that the stochastic gradients are bounded almost surely. The latter is a very restrictive assumption that is not fulfilled even by Gaussian noise. In [66] a clipped algorithm with momentum is shown to converge in expectation to a stationary point under the same strong assumptions on the noise. Similar assumptions are also encountered in e.g. [13, 34]; the latter also considers the slightly more general case of *sub-Gaussian* noise. [29] analyses clipped SGD under Assumption 4.ii), but do not obtain a convergence guarantee due to an *unavoidable bias* [29]. Recently, [56] and [17] obtained convergence guarantees for versions of AdaGradNorm under the weaker affine variance-assumption. These results are however only with a certain probability, and there is always some set of positive measure on which the algorithm may not converge. The convergence guarantees that we obtain in Theorem 3.5 under Assumption 4.i) and Assumption 3.ii) is stronger in the sense that it converges for almost every path. We also stress the fact that Assumption 3.ii) is relatively weak since it covers all heavy-tailed distributions with finite variance [49]. This includes for instance the large class of sub-Weibull distributions, which generalizes sub-Gaussian and sub-exponential distributions [55].

3. Analysis

We first give a brief overview of the analysis in Section 3.1. In Section 3.2 we describe the setting and in Section 3.3 we give a more detailed outline of the theorems and the proofs. The full proofs of the results are given in Appendix A.

3.1. Brief overview

The analysis is split into two parts.

In the first, we show that the iterates of the scheme defined by (10) are finite almost surely, if the objective function F and the convex kinetic energy function φ are L-smooth and coercive, or if F is (L_0, L_1)-smooth and the variance is finite. This is done by constructing a Lyapunov function with the help of the Hamiltonian H, and then appealing to the classical Robbins–Siegmund theorem [48].

In the second, we show that given that the iterates defined by (10) are bounded, they converge almost surely to a stationary point of F. We make use of a modification of the *ODE method*, compare [31], in conjunction with techniques from [7]. Since the scheme is implicit-explicit, we cannot directly apply e.g. Theorem 2.1 in [31]. Essentially, the idea is to

i) Introduce a pseudo-time $t_k = \sum_{i=0}^{k-1} \alpha_i$ and construct piecewise constant interpolations $P_0(t)$ and $Q_0(t)$ of $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$ from (10).

ii) Show that the time-shifted processes $P_k(t) = P_0(t_k + t)$ and $Q_k(t) = Q_0(t_k + t)$ are *equicontinuous in the extended sense* [31] and that (P_0, Q_0) is an *asymptotic pseudotrajectory* [7] for (9). This essentially means that (P_k, Q_k) asymptotically satisfies (9).

iii) Use the previous point to show that the set of limit points of $\{(p_k, q_k)\}_{k\geq 0}$ is *internally chain transitive* [8] for the flow generated by (9).

iv) At last, make use of the underlying Hamiltonian dynamics of (9) to conclude that all limit points of $\{(p_k, q_k)\}_{k\geq 0}$ are equilibria of the flow generated by (9). Thus $\{q_k\}_{k\geq 0}$ converges to a stationary point of F.

All the conclusions in the above four points should be interpreted as holding almost surely.

3.2. Setting

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\{\xi_k\}_{k\geq 0}$ be a sequence of independent, identically distributed random variables. We denote a generic outcome in the underlying sample space Ω by ω. We further let $\mathcal{F}_k$ denote the σ-algebra generated by $\xi_0, \dots, \xi_{k-1}$. By $\mathbb{E}_{\xi_k}[X]$ we denote the conditional expectation of a random variable X with respect to $\mathcal{F}_k$. For a set $A \subset \mathbb{R}^d$, we let $N_\delta(A) = \{x : \inf_{a\in A} \|x - a\| < \delta\}$. Here, $\|\cdot\|$ denotes the standard Euclidean norm on $\mathbb{R}^d$. In the rest of the paper, we will also use it to denote the Euclidean norm on $\mathbb{R}^{2d}$, since it should be clear from the context whether the argument is in $\mathbb{R}^d$ or $\mathbb{R}^{2d}$.

Throughout, we denote by ϕ the semiflow generated by (9). The set of limit points of ϕ is denoted by

$$\mathrm{Lim}(\phi) = \{z : \exists x \in \mathbb{R}^{2d} \text{ and } \{t_k\}_{k\geq 0} \text{ with } \lim_{k\to\infty} t_k = \infty \text{ for which } \phi(t_k, x) \to z\},$$

Likewise, the set of limit points of a function Z will be denoted by

$$\mathrm{Lim}(Z) = \{z : \exists \{t_k\}_{k\geq 0} \text{ with } \lim_{k\to\infty} t_k = \infty \text{ for which } Z(t_k) \to z\},$$

and we use the same notation for the limit points of a sequence $\{z_k\}_{k\geq 0}$, i.e.

$$\mathrm{Lim}(\{z_k\}_{k\geq 0}) = \{z : \text{for any } \epsilon > 0, \exists \{n_k\}_{k\geq 0} \text{ such that } \|z_{n_k} - z\| \leq \epsilon\}.$$

Finally, the set of equilibria of (9) is denoted by

$$E_H = \{(p, q) \,:\, \nabla H(p, q) = 0\}.$$

3.2.1. *Basic assumptions*

We make the following basic assumptions on F, f and φ:

Assumption 1. The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable and satisfies:

i) *(Coercivity)* $\lim_{\|x\|\to\infty} F(x) = \infty$.

ii) *(Empty interior)* Let $\Lambda = \{q : \nabla F(q) = 0\}$. The set $F(\Lambda)$ has empty interior.

Further, the stochastic gradient ∇f is an unbiased estimator of ∇F, i.e.

iii) $\mathbb{E}[\nabla f(x, \xi)] = \nabla F(x)$.

The infimum of F, whose existence is implied by the continuity of F and i), is denoted by F_*. It satisfies $F(x) \geq F_* > -\infty$ for all $x \in \mathbb{R}^d$.

Remark 3.1. Assumption 1.ii) is needed in the proof of Proposition A.11 (Proposition 6.4 in [8]). The proof of the proposition makes use of the fact that any point in the image of the critical set can be approximated by points outside of the set (see the proof of Proposition 6.4 in [8]). Assumption 1.ii) is for instance satisfied for any $F \in C^\infty$ by Sard's theorem, compare [50]. An example of a function that violates Assumption 1.ii) was given by Whitney in [58], where a C^1-function that is non-constant on a connected set of critical points is constructed.

Assumption 2. The kinetic energy function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is differentiable and satisfies:

i) *(Lipschitz continuous $\nabla\varphi$)* There is a constant $\lambda > 0$ such that $\|\nabla\varphi(y) - \nabla\varphi(x)\| \leq \lambda\|x - y\|$, for all $x, y \in \mathbb{R}^d$.
ii) *(Convexity)* For all $x, y \in \mathbb{R}^d$, it holds that $\varphi(y) - \varphi(x) \leq \langle \nabla\varphi(y), y - x\rangle$.
iii) *(Coercivity)* $\lim_{\|x\|\to\infty} \varphi(x) = \infty$.

The infimum of φ, whose existence is implied by the continuity of φ and iii), is denoted by φ_*. It satisfies $\varphi(x) \geq \varphi_* > -\infty$ for all $x \in \mathbb{R}^d$.

Remark 3.2. Assumption 1.i) and 2.iii) together implies that the Hamiltonian $H(p, q) = F(q)+\varphi(p)$ is coercive as a function of q and p.

In addition to these basic assumptions, we consider three different settings.

3.2.2. *Setting 1*

In the first setting, ∇F is Lipschitz continuous but the stochastic gradients can have large variance.

Assumption 3. The objective function F and the stochastic gradient ∇f further satisfy:

i) *(Lipschitz-continuous ∇F)* There is a constant $L > 0$ such that $\|\nabla F(y) - \nabla F(x)\| \leq L\|x - y\|$, for all $x, y \in \mathbb{R}^d$.
ii) *(Locally bounded variance)* $\mathbb{V}\left[\nabla f(x, \xi)\right] \leq \kappa\left(F(x) - F_*\right) + \tau\|\nabla F(x)\|^2 + \sigma^2$, where $\kappa, \sigma, \tau \geq 0$.

Here $\mathbb{V}\left[\nabla f(x, \xi)\right] = \mathbb{E}\left[\|\nabla f(x, \xi)\|^2\right] - \|\nabla F(x)\|^2 = \mathbb{E}\left[\|\nabla f(x, \xi) - \nabla F(x)\|^2\right]$.

Assumption 3.i) implies that the inequality $F(y) - F(x) \leq \langle \nabla F(x), y - x\rangle + \frac{L}{2}\|x - y\|^2$ holds for all $x, y \in \mathbb{R}^d$, compare Lemma 1.2.3 in [40]. Assumption 3.ii) was first introduced in [28] where it was called "expected smoothness". It is weak in the sense that it allows for infinite variance in the case that either the gradient or the objective function becomes infinitely large, see [28, Section 4.5] for several settings in which it is satisfied. The condition is similar to e.g. the "affine noise variance" in [57] and the "affine variance" in [17].

3.2.3. *Setting 2*

Alternatively, we consider the following setting, where we require less regularity of F but instead restrict the variance of the stochastic gradient.

Assumption 4. The objective function F and the stochastic gradient ∇f further satisfy:

i) *((L_0, L_1)-smoothness)* There exists L_0, L_1 such that for all $x, y \in \mathbb{R}^d$, if $\|x - y\| \leq \frac{1}{L_1}$, then
$$\|\nabla F(x) - \nabla F(y)\| \leq \left(L_0 + L_1\|\nabla F(y)\|\right)\|x - y\|.$$
ii) *(Bounded variance)* $\mathbb{V}\left[\nabla f(x, \xi)\right] \leq \sigma^2$,
iii) *(Bounded $\nabla\varphi$)* There exists $\Delta > 0$ such that $\|\nabla\varphi(x)\| \leq \Delta$ for all $x \in \mathbb{R}$.

Remark 3.3. By *Lyapunov's inequality*, $\mathbb{E}[X^2] \leq \mathbb{E}[X]^{1/2}$ (cf. [53, p. 230]), it follows from Assumption 4.ii) that
$$\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla F(x)\|\right] \leq \sigma.$$

Assumption 4.ii) is sometimes referred to as a *heavy-tailed noise* assumption in the literature [25, 29]. It covers all zero-mean, heavy-tailed distributions with finite second moment, compare [49]. In particular, it also includes the large class of *sub-Weibull distributions* [55], which generalizes random variables of sub-Gaussian and sub-Exponential distribution.

3.2.4. *Setting 3*

Assumption 4.ii) may be restrictive in some cases, compare [26]. Therefore we also consider the empirical risk minimization setting when the objective function is on the form (2), $f(\cdot,\xi)$ is (L_0, L_1)-smooth and $\nabla f(\cdot,\xi)$ is given by (4). In this setting, we can further lower the assumptions on the noise to merely finite expectation:

Assumption 5. The objective function satisfies 4.i) and $\nabla\varphi$ satisfies 4.iii). The objective function F and the stochastic gradient ∇f further satisfy:

- i) *(Empirical risk minimization)* The objective function and the stochastic gradient are of the form (2) and (4), where for each i it holds that $\inf_{q\in\mathbb{R}^d} f_i(q) > -\infty$.
- ii) *(Bounded expectation)* The stochastic gradients satisfy $\mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|\right] \le \sigma$.
- iii) *((L_0, L_1)-smooth)* The stochastic functions $f(\cdot,\xi)$ are (L_0, L_1)-smooth.

Remark 3.4. We observe that SGD with momentum, corresponding to (5), requires $\varphi(x) = \|x\|^2/2$, which does not satisfy Assumption 4.iii). It is thus only covered by the analysis in Setting 1. In Setting 2 and 3, we have a weaker regularity assumption on F, and this requires us to instead pose stricter requirements on the methods. Overall this indicates that for (L_0, L_1)-smooth cost functionals, clipping methods are a better option than SGD.

3.2.5. *Book-keeping assumptions*

The following assumption on the step sizes α_k is standard and originates from [47]. Informally, the step sizes must go to zero in order to counter the stochasticity, but do so slowly enough that we have time to reach a stationary point.

Assumption 6. The step size sequence $\{\alpha_k\}_{k\ge0}$ satisfies $\alpha_0 = 0$ and $\{\alpha_k\}_{k\ge0} \in \ell^2(\mathbb{R})\backslash\ell^1(\mathbb{R})$.

Our analysis shows convergence to the set of stationary points of ∇F. Under the following additional assumption, we get convergence to a unique stationary point:

Assumption 7. The stationary points of F are isolated.

3.3. **Outline of analysis**

The proofs of the results in this section can be found in Appendix A. The main theorem is as follows:

Theorem 3.5. *Let Assumptions 1, 2 and 6 be satisfied, as well as either Assumption 3, 4 or 5. Then $\{q_k\}_{k\ge0}$ converges almost surely to the set of stationary points of the objective function F. If we additionally assume that Assumption 7 holds, the convergence is to a unique stationary point.*

The following result is a direct consequence of Theorem 3.5:

Corollary 3.6 (Convergence in expectation)**.** *Let Assumptions 1, 2, 6 and 7 be valid. Further, let the Hamiltonian be on the form (7) and let the sequences $\{p_k\}_{k\ge0}$ and $\{q_k\}_{k\ge0}$ be generated by (10). Then it holds that*

$$\lim_{k\to\infty} \mathbb{E}\left[\|\nabla F(q_k)\|^\theta\right] = 0,$$

where $0 < \theta < 2$ under Assumption 3 and $0 < \theta < 1$ under Assumption 4 or 5.

Our proof strategy consists of two parts. In the first part we show that the sequences $\{p_k\}_{k\ge0}$ and $\{q_k\}_{k\ge0}$ are finite almost surely:

Theorem 3.7 (Finiteness of $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$)**.** *Let Assumptions 1, 2 and 6 be valid, as well as either Assumption 3, 4 or 5. Further, let the Hamiltonian be on the form (7) and let the sequences $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$ be generated by (10). Then $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$ are finite almost surely. Moreover, it holds that $\sup_{k\geq 0}\mathbb{E}\left[F(q_k)\right]<\infty$.*

In the second part of the analysis, we closely follow the ODE method approach as outlined in [31]: We start by introducing a pseudo-time $t_k=\sum_{i=0}^{k-1}\alpha_i$, and define two piecewise constant, (stochastic) interpolation processes by

$$\begin{aligned} P_0(t) &= p_0 I_{(-\infty,t_0]}(t)+\sum_{k=0}^{\infty} p_k I_{[t_k,t_{k+1})}(t),\\ Q_0(t) &= q_0 I_{(-\infty,t_0]}(t)+\sum_{k=0}^{\infty} p_k I_{[t_k,t_{k+1})}(t). \end{aligned} \tag{12}$$

We next consider the shifted sequence of processes $\{P_k\}_{k\geq 0}$ and $\{Q_k\}_{k\geq 0}$, defined by

$$\begin{aligned} P_k(t) &= P_0(t_k+t),\\ Q_k(t) &= Q_0(t_k+t). \end{aligned} \tag{13}$$

We note that $\{P_k\}_{k\geq 0}$ and $\{Q_k\}_{k\geq 0}$ are stochastic processes; they depend on $\omega\in\Omega$ through the stochasticity of the sequences $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$. For brevity we will refrain from writing out the dependence on ω.

The next step is to introduce the concept of *extended equicontinuity* [20,31]:

Definition 3.8 (Extended equicontinuity)**.** A sequence of $\mathbb{R}^d$-valued functions $\{f_k\}_{k\geq 0}$, defined on $(-\infty,\infty)$, is said to be *equicontinuous in the extended sense* if $\{\|f_k(0)\|\}_{k\geq 0}$ is bounded and for every $T>0$ and $\epsilon>0$ there is a $\delta>0$ such that

$$\limsup_{k\to\infty}\sup_{0<|t-s|\leq\delta,\ t,s\in[0,T]}\|f_k(t)-f_k(s)\|\leq\epsilon. \tag{14}$$

Following [20], we show that the process $\{Z_k\}_{k\geq 0}=\{(P_k,Q_k)\}_{k\geq 0}$ is equicontinuous in the extended sense:

Lemma 3.9 (Equicontinuous in the extended sense)**.** *Consider $\{Z_k\}_{k\geq 0}=\{(P_k,Q_k)\}_{k\geq 0}$ where the sequences $\{P_k\}_{k\geq 0}$ and $\{Q_k\}_{k\geq 0}$ are defined by (13) (equivalently, by (30)). Suppose that $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$ are defined by (10), and that the Hamiltonian is on the form (7). Further, let Assumptions 1, 2 and 6 be valid, as well as either Assumption 3, 4 or 5. Then $\{Z_k\}_{k\geq 0}$ is equicontinuous in the extended sense, almost surely.*

This property is used to show that Z_0 is an *asymptotic pseudotrajectory* [8] to (9) in the following sense:

Lemma 3.10 (Asymptotic pseudotrajectory)**.** *Let Assumptions 1, 2 and 6 be valid, as well as either Assumption 3, 4 or 5. Then the function $Z_0=(P_0,Q_0)$ where P_0 and Q_0 are defined by (12) is almost surely an asymptotic pseudotrajectory to the semiflow ϕ generated by equation (9). That is,*

$$\lim_{s\to\infty}\sup_{t\in[0,T]}\|Z_0(s+t)-\phi(t,Z_0(s))\|=0,$$

for all $T>0$.

The main consequence of this is that the set $\mathrm{Lim}(\{(p_k,q_k)\}_{k\geq 0})$ is *internally chain transitive*. The following proposition is essentially a paraphrasing of [7, Corollary 4.3] combined with [8, Proposition 5.3]:

Lemma 3.11 (Internally chain transitive)**.** *Let Assumptions 1, 2 and 6 be valid, as well as either Assumption 3, 4 or 5. Then given that $\{z_k\}_{k\geq 0}=\{(p_k,q_k)\}_{k\geq 0}$ is determined by (10), the set $\mathrm{Lim}(\{z_k\}_{k\geq 0})$ is a compact, and internally chain transitive set for the semiflow ϕ generated by (9).*

The final step uses the Hamiltonian dynamics. We have

Lemma 3.12 (Lyapunov function). *Let Assumptions 1 and 2 be valid, and let $V : \mathbb{R}^{2d} \to \mathbb{R}$ be defined by*

$$V(z) = H(p, q) - F_* - \varphi_*, \tag{15}$$

where $z = (p, q)$. Then V is a Lyapunov function for $E_H \subset \mathbb{R}^{2d}$ with respect to ϕ.

It follows from Lemma 3.12 that V is also a Lyapunov function for $\Lambda = \mathrm{Lim}(\{z_k\}_{k\geq 0}) \cap E_H$ with respect to the restricted flow $\Psi = \phi|_{\mathrm{Lim}(\{z_k\}_{k\geq 0})}$, and it can be shown that $\Lambda \neq \emptyset$. Making use of Assumption 1.ii), we can appeal to [8, Proposition 6.4] (see Remark 3.1) to conclude that any internally chain transitive set for ϕ is a subset of Λ. Since $\mathrm{Lim}(\{z_k\}_{k\geq 0})$ is such a set by Lemma 3.11, we get $\mathrm{Lim}(\{z_k\}_{k\geq 0}) \subset E_H$. This proves the main assertion of the paper.

4. Conclusions

In this paper, we have shown that the stochastic Hamiltonian descent algorithm (10), arising as a stochastic implicit-explicit Euler discretization of (9), under weak assumptions converges almost surely to the set of stationary points of the objective function F. In the terminology of [47], this means that the estimator determined by $\{q_k\}$ is a strongly consistent estimator of a stationary point of F. (Here, the designated asymptoticity is with respect to the number of iterations instead of the sample size.) Similarly, the result in Corollary 3.6 is akin to $\{q_k\}_{k\geq 0}$ being an asymptotically unbiased estimator of a stationary point q_*.

5. Acknowledgements

The authors are grateful to the anonymous reviewers, whose constructive feedback led to a significantly shorter and much improved manuscript.

Appendix A. Analysis

As explained in Section 3.3, the two main steps of the convergence analysis are to first prove that p_k and q_k are finite almost surely, and then to use this a priori result to show that they in fact converge.

A.1. The sequences $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$ are finite almost surely

We first prove Theorem 3.7:

Theorem 3.7 (Finiteness of $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$). *Let Assumptions 1, 2 and 6 be valid, as well as either Assumption 3, 4 or 5. Further, let the Hamiltonian be on the form (7) and let the sequences $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$ be generated by (10). Then $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$ are finite almost surely. Moreover, it holds that $\sup_{k\geq 0} \mathbb{E}\left[F(q_k)\right] < \infty$.*

The proof relies on the Robbins-Siegmund theorem:

Theorem A.1 ([48]). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$ be a sequence of sub-σ-algebras of $\mathcal{F}$. For each $k = 1, 2, \dots$ let V_k, β_k, X_k and Y_k be non-negative $\mathcal{F}_k$-measurable random variables such that*

$$\mathbb{E}\left[V_{k+1}|\mathcal{F}_k\right] \leq V_k(1 + \beta_k) + X_k - Y_k.$$

Then on the set $\left\{\omega : \sum_k \beta_k < \infty, \sum_k X_k < \infty\right\}$, the limit

$$\lim_{k\to\infty} V_k = V$$

exists and is finite and $\sum_k Y_k < \infty$.

We first consider Setting 1, i.e. with Assumption 3. The strategy is to introduce $V_k = H(p_k, q_k) - F_* - \varphi_* = F(q_k) - F_* + \varphi(p_k) - \varphi_*$, then use L-smoothness of F and the convexity of φ to bound the difference $V_{k+1} - V_k$ by $\alpha_k^2 V_k$ plus higher-order terms of α_k. Then we can appeal to Theorem A.1 to conclude that $\{p_k\}_{k\geq 1}$ and $\{q_k\}_{k\geq 1}$ are finite a.s.

Proof of Theorem 3.7 in Setting 1. Let $V_k = H(p_k, q_k) - F_* - \varphi_*$. Then we have that

$$V_{k+1} - V_k = F(q_{k+1}) - F(q_k) + \varphi(p_{k+1}) - \varphi(p_k). \tag{16}$$

By L-smoothness of F and convexity of φ, this is less than or equal to

$$\langle \nabla F(q_k), q_{k+1} - q_k \rangle + \frac{L}{2}\|q_{k+1} - q_k\|^2 + \langle \nabla\varphi(p_{k+1}), p_{k+1} - p_k \rangle.$$

We insert (10) into the previous expression to obtain that it is equal to

$$I_1 + I_2 + I_3 := \alpha_k \langle \nabla F(q_k) - \nabla f(q_k, \xi_k), \nabla\varphi(p_{k+1})\rangle + \frac{L\alpha_k^2}{2}\|\nabla\varphi(p_{k+1})\|^2 - \alpha_k\gamma\langle \nabla\varphi(p_{k+1}), \nabla\varphi(p_k)\rangle,$$

where

$$I_1 = \alpha_k \langle \nabla F(q_k) - \nabla f(q_k, \xi_k), \nabla\varphi(p_k)\rangle + \alpha_k \langle \nabla F(q_k) - \nabla f(q_k, \xi_k), \nabla\varphi(p_{k+1}) - \nabla\varphi(p_k)\rangle.$$

When we take the conditional expectation (with respect to the sigma algebra generated by $\xi_1, \dots, \xi_{k-1}$) of I_1, the first term is 0 by the unbiasedness of the gradient and the independence of $\{\xi_k\}$:

$$\mathbb{E}_{\xi_k}[I_1] = \alpha_k \mathbb{E}_{\xi_k}\left[\langle \nabla F(q_k) - \nabla f(q_k, \xi_k), \nabla\varphi(p_{k+1}) - \nabla\varphi(p_k)\rangle\right].$$

Using Assumption 2.i), we can bound I_3 as

$$I_3 = -\alpha_k\gamma\langle \nabla\varphi(p_{k+1}), \nabla\varphi(p_k)\rangle \leq \frac{\alpha_k\gamma}{2}\|\nabla\varphi(p_{k+1}) - \nabla\varphi(p_k)\|^2 \leq \frac{\alpha_k\gamma\lambda^2}{2}\|p_{k+1} - p_k\|^2.$$

After taking the expectation of (16), we thus get the bound

$$\begin{aligned} \mathbb{E}_{\xi_k}[V_{k+1}] - V_k &\leq \alpha_k \mathbb{E}_{\xi_k}\left[\langle \nabla F(q_k) - \nabla f(q_k, \xi_k), \nabla\varphi(p_{k+1}) - \nabla\varphi(p_k)\rangle\right] \\ &\quad + \frac{L\alpha_k^2}{2}\mathbb{E}_{\xi_k}\left[\|\nabla\varphi(p_{k+1})\|^2\right] + \frac{\alpha_k\gamma\lambda^2}{2}\mathbb{E}_{\xi_k}\left[\|p_{k+1} - p_k\|^2\right] \\ &=: I_1' + I_2' + I_3'. \end{aligned} \tag{17}$$

We now make use of Cauchy–Schwarz inequality along with the Lipschitz continuity of $\nabla\varphi$ to bound I_1' as

$$I_1' \leq \alpha_k\lambda \mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k, \xi_k)\|\|p_{k+1} - p_k\|\right].$$

We insert (10) into the previous expression, and make use of Young's inequality for products, $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$, to obtain that

$$\begin{aligned} I_1' &\leq \alpha_k^2\lambda \mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k, \xi_k)\|\|\nabla f(q_k, \xi_k) + \gamma\nabla\varphi(p_k)\|\right] \\ &\leq \frac{\alpha_k^2}{2}\lambda \mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k, \xi_k)\|^2\right] + \frac{\alpha_k^2}{2}\lambda \mathbb{E}_{\xi_k}\left[\|\nabla f(q_k, \xi_k) + \gamma\nabla\varphi(p_k)\|^2\right]. \end{aligned}$$

Making use of the inequality

$$\|x - y\|^2 \leq 2\|x\|^2 + 2\|y\|^2, \tag{18}$$

we can further bound I_1' by

$$I_1' \leq \frac{\alpha_k^2}{2}\lambda\mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k,\xi_k)\|^2\right] + \alpha_k^2\lambda\mathbb{E}_{\xi_k}\left[\|\nabla f(q_k,\xi_k)\|^2\right] + \alpha_k^2\lambda\gamma^2\mathbb{E}_{\xi_k}\left[\|\nabla\varphi(p_k)\|^2\right].$$

At last we make use of Assumption 3.ii) to get that

$$\begin{aligned} I_1' &\leq \frac{\alpha_k^2}{2}\lambda\left(\kappa(F(q_k)-F_*) + \tau\|\nabla F(q_k)\|^2 + \sigma^2\right) \\ &\quad + \alpha_k^2\lambda\left(\kappa(F(q_k)-F_*) + (1+\tau)\|\nabla F(q_k)\|^2 + \sigma^2\right) + \alpha_k^2\lambda\gamma^2\mathbb{E}_{\xi_k}\left[\|\nabla\varphi(p_k)\|^2\right]. \end{aligned}$$

We now turn our attention to the term I_2' in (17). Adding and subtracting $\nabla\varphi(p_k)$ and making use of Assumption 2.i) we get that

$$\begin{aligned} I_2' &\leq L\alpha_k^2\mathbb{E}_{\xi_k}\left[\|\nabla\varphi(p_{k+1}) - \nabla\varphi(p_k)\|^2\right] + L\alpha_k^2\mathbb{E}_{\xi_k}\left[\|\nabla\varphi(p_k)\|^2\right] \\ &\leq L\alpha_k^2\lambda^2\mathbb{E}_{\xi_k}\left[\|p_{k+1}-p_k\|^2\right] + L\alpha_k^2\|\nabla\varphi(p_k)\|^2 \\ &\leq 2L\alpha_k^4\lambda^2\mathbb{E}_{\xi_k}\left[\|\nabla f(q_k,\xi_k)\|^2\right] + \left(2L\alpha_k^4\gamma^2\lambda^2 + L\alpha_k^2\right)\|\nabla\varphi(p_k)\|^2, \end{aligned}$$

where we have used (10) and (18) in the last step. Making use of Assumption 3.ii) again we obtain that

$$I_2' \leq 2L\alpha_k^4\lambda^2\left(\kappa(F(q_k)-F_*) + (1+\tau)\|\nabla F(q_k)\|^2 + \sigma^2\right) + \left(2L\alpha_k^4\gamma^2\lambda^2 + L\alpha_k^2\right)\|\nabla\varphi(p_k)\|^2.$$

In a similar way, we find that

$$I_3' \leq \alpha_k^3\gamma\lambda^2\left(\kappa(F(q_k)-F_*) + (1+\tau)\|\nabla F(q_k)\|^2 + \sigma^2\right) + \alpha_k^3\gamma^3\lambda^2\|\nabla\varphi(p_k)\|^2.$$

Gathering up the terms, we get that

$$\begin{aligned} \mathbb{E}_{\xi_k}[V_{k+1}] - V_k &\leq \left(\frac{\alpha_k^2\lambda}{2} + \alpha_k^2\lambda + 2L\alpha_k^4\lambda^2 + \alpha_k^3\gamma\lambda^2\right)\sigma^2 \\ &\quad + \alpha_k^2\kappa\lambda\left(\frac{3}{2} + 2L\alpha_k^2\lambda + \alpha_k\gamma\lambda\right)(F(q_k)-F_*) \\ &\quad + \left(\frac{\alpha_k^2\lambda\tau}{2} + \alpha_k^2\lambda(1+\tau) + 2L\alpha_k^4\lambda^2(1+\tau) + \alpha_k^3\gamma\lambda^2(1+\tau)\right)\|\nabla F(q_k)\|^2 \\ &\quad + \left(\alpha_k^2\gamma^2\lambda + 2L\alpha_k^4\lambda^2\gamma^2 + L\alpha_k^2 + \alpha_k^3\gamma^3\lambda^2\right)\|\nabla\varphi(p_k)\|^2. \end{aligned}$$

Making use of Lemma B.1, we see that

$$\begin{aligned} \mathbb{E}_{\xi_k}[V_{k+1}] - V_k &\leq \left(\frac{\alpha_k^2\lambda}{2} + \alpha_k^2\lambda + 2L\alpha_k^4\lambda^2 + \alpha_k^3\gamma\lambda^2\right)\sigma^2 \\ &\quad + (F(q_k)-F_*)\Bigg[\alpha_k^2\kappa\lambda\left(\frac{3}{2} + 2L\alpha_k^2\lambda + \alpha_k\gamma\lambda\right) \\ &\qquad + 2L\left(\frac{\alpha_k^2\lambda\tau}{2} + \alpha_k^2\lambda(1+\tau) + 2L\alpha_k^4\lambda^2(1+\tau) + \alpha_k^3\gamma\lambda^2(1+\tau)\right)\Bigg] \\ &\quad + \left(\alpha_k^2\gamma^2\lambda + 2L\alpha_k^4\lambda^2\gamma^2 + L\alpha_k^2 + \alpha_k^3\gamma^3\lambda^2\right)(\varphi(p_k)-\varphi_*). \end{aligned}$$

Now define

$$C_1(\alpha_k) = \sigma^2 \left(\frac{\alpha_k^2 \lambda}{2} + \alpha_k^2 \lambda + 2L\alpha_k^4 \lambda^2 + \alpha_k^3 \gamma \lambda^2 \right)$$

and let $C_2(\alpha_k)$ be the maximum of the terms multiplying $F(q_k) - F_*$ and $\varphi(p_k) - \varphi_*$. It follows that

$$\mathbb{E}_{\xi_k}\left[V_{k+1}\right] - V_k \leq C_1(\alpha_k) + C_2(\alpha_k) V_k. \tag{19}$$

Since C_1 and C_2 only contain second-order terms of α_k (and by assumption $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$), we have that

$$\sum_{k=0}^{\infty} C_1(\alpha_k) < \infty, \ \sum_{k=0}^{\infty} C_2(\alpha_k) < \infty.$$

We can thus make use of the Robbins–Siegmund theorem A.1 with $\beta_k = C_2(\alpha_k)$, $X_k = C_1(\alpha_k)$ and $Y_k = 0$ to conclude that V_k tends to a non-negative, finite, random variable V almost surely. Since F and φ are assumed to be coercive, this implies that $\{p_k\}_{k\geq 1}$ and $\{q_k\}_{k\geq 1}$ are finite almost surely. For the second claim of the proof, we define

$$S_k = \frac{V_k}{\prod_{j=0}^{k-1}(1 + C_2(\alpha_j))}.$$

By (19), we have that

$$\mathbb{E}_{\xi_k}\left[S_{k+1}\right] \leq S_k + \frac{C_1(\alpha_k)}{\prod_{j=0}^{k}(1 + C_2(\alpha_j))} \leq S_k + C_1(\alpha_k).$$

Taking the expectation and, summing from 0 to $K-1$, we see that

$$\mathbb{E}\left[S_K\right] \leq S_0 + \sum_{k=0}^{K-1} C_1(\alpha_k).$$

We multiply both sides of the previous inequality with $\prod_{k=0}^{K-1}(1 + C_2(\alpha_k))$

$$\begin{aligned}
\mathbb{E}\left[V_K\right] &\leq \prod_{k=0}^{K-1}(1 + C_2(\alpha_k)) \left(S_0 + \sum_{k=0}^{K-1} C_1(\alpha_k) \right) \\
&\leq e^{\sum_{j=0}^{K-1} C_2(\alpha_k))} \cdot \left(S_0 + \sum_{k=0}^{K-1} C_1(\alpha_k) \right),
\end{aligned}$$

where we have used the fact that $1 + x \leq e^x$ in the second step. Letting K tend to infinity on the left hand side and using the fact that $\sum_{k=0}^{\infty} C_2(\alpha_k)) < \infty$ we see that the last claim of the theorem also holds:

$$\sup_k \mathbb{E}\left[V_k\right] < \infty.$$

□

We now give a proof of Theorem 3.7 in Setting 2, i.e. with Assumption 4.

Proof of Theorem 3.7 in Setting 2. By Assumption 4.i) it holds for $\|x_k - x_{k+1}\| \leq \frac{1}{L_1}$ that

$$F(x_{k+1}) - F(x_k) \leq \langle \nabla F(x_k), x_{k+1} - x_k \rangle + \frac{L_0 + L_1\|\nabla F(x_k)\|}{2}\|x_{k+1} - x_k\|^2. \tag{20}$$

Since

$$q_{k+1} - q_k = \alpha_k \nabla\varphi(p_{k+1}) \tag{21}$$

we get for large enough k that

$$\|q_{k+1} - q_k\| = \alpha_k \|\nabla\varphi(p_{k+1})\| \leq \alpha_k \Delta \leq \frac{1}{L_1}, \tag{22}$$

by Assumption 4.iii). If we insert (21) into (20), we get

$$F(q_{k+1}) - F(q_k) \leq \langle \nabla F(q_k), q_{k+1} - q_k\rangle + \frac{L_0}{2}\|q_{k+1} - q_k\|^2 + \frac{L_1}{2}\|\nabla F(q_k)\|\|q_{k+1} - q_k\|^2.$$

By (22), we get that

$$F(q_{k+1}) - F(q_k) \leq \alpha_k\langle \nabla F(q_k), \nabla\varphi(p_{k+1})\rangle + \frac{L_0}{2}\alpha_k^2\Delta^2 + \frac{L_1}{2}\alpha_k^2\Delta^2\|\nabla F(q_k)\|.$$

By Assumption 2.ii) we have that

$$\begin{aligned}\varphi(p_{k+1}) - \varphi(p_k) &\leq \langle \nabla\varphi(p_{k+1}), p_{k+1} - p_k\rangle \\ &= -\alpha_k\langle \nabla\varphi(p_{k+1}), \nabla f(q_k, \xi_k)\rangle - \alpha_k\gamma\langle \nabla\varphi(p_{k+1}), \nabla\varphi(p_k)\rangle\end{aligned}$$

With $H(p, q) = F(q) + \varphi(p)$ and $V_k = H(p_k, q_k) - F_* - \varphi_*$ as in the previous proof we thus get that

$$\begin{aligned}V_{k+1} - V_k \leq\ &\alpha_k\langle \nabla F(q_k), \nabla\varphi(p_{k+1})\rangle + \frac{L_0}{2}\alpha_k^2\Delta^2 + \frac{\alpha_k^2 L_1\Delta^2}{2}\|\nabla F(q_k)\| \\ &- \alpha_k\langle \nabla\varphi(p_{k+1}), \nabla f(q_k, \xi_k)\rangle - \alpha_k\gamma\langle \nabla\varphi(p_{k+1}), \nabla\varphi(p_k)\rangle,\end{aligned}$$

which can be rewritten as

$$\begin{aligned}V_{k+1} - V_k \leq\ &\alpha_k\langle \nabla F(q_k) - \nabla f(q_k, \xi_k), \nabla\varphi(p_{k+1})\rangle + \frac{L_0}{2}\alpha_k^2\Delta^2 + \frac{\alpha_k^2 L_1\Delta^2}{2}\|\nabla F(q_k)\| \\ &- \alpha_k\gamma\langle \nabla\varphi(p_{k+1}), \nabla\varphi(p_k)\rangle.\end{aligned}$$

We add and subtract $\nabla\varphi(p_k)$ in the first scalar product:

$$\begin{aligned}V_{k+1} - V_k \leq\ &\alpha_k\langle \nabla F(q_k) - \nabla f(q_k, \xi_k), \nabla\varphi(p_{k+1}) - \nabla\varphi(p_k)\rangle \\ &+ \alpha_k\langle \nabla F(q_k) - \nabla f(q_k, \xi_k), \nabla\varphi(p_k)\rangle \\ &+ \frac{L_0}{2}\alpha_k^2\Delta^2 + \frac{\alpha_k^2 L_1\Delta^2}{2}\|\nabla F(q_k)\| - \alpha_k\gamma\langle \nabla\varphi(p_{k+1}), \nabla\varphi(p_k)\rangle \\ =:\ &I_1 + I_2 + I_3.\end{aligned} \tag{23}$$

Due to the unbiasedness of $\nabla f(q_k, \xi_k)$ and the fact that ξ_k is independent of q_k and p_k, we find that $\mathbb{E}_{\xi_k}[I_2] = 0$. We now focus on I_1. Taking the conditional expectation and using Cauchy–Schwarz inequality and the Lipschitz continuity of $\nabla\varphi$, we see that

$$\mathbb{E}_{\xi_k}[I_1] \leq \alpha_k^2\lambda\mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k, \xi_k)\|\left(\|\nabla f(q_k, \xi_k)\| + \gamma\|\nabla\varphi(p_k)\|\right)\right]. \tag{24}$$

We then add and subtract $\nabla F(q_k)$ inside the $\|\nabla f(q_k,\xi_k)\|$-term and use the triangle inequality as well as the assumption that $\|\nabla\varphi(p_k)\| \le \Delta$:

$$\begin{aligned}
\mathbb{E}_{\xi_k}[I_1] &\le \alpha_k^2\lambda\mathbb{E}_{\xi_k}\left[\|\nabla F(q_k)-\nabla f(q_k,\xi_k)\|\left(\|\nabla F(q_k)-\nabla f(q_k,\xi_k)\|+\|\nabla F(q_k)\|+\gamma\Delta\right)\right]\\
&= \alpha_k^2\lambda\mathbb{E}_{\xi_k}\left[\|\nabla F(q_k)-\nabla f(q_k,\xi_k)\|^2+\|\nabla F(q_k)-\nabla f(q_k,\xi_k)\|\|\nabla F(q_k)\|+\gamma\Delta\|\nabla F(q_k)-\nabla f(q_k,\xi_k)\|\right]\\
&= \alpha_k^2\lambda\Big(\mathbb{E}_{\xi_k}\left[\|\nabla F(q_k)-\nabla f(q_k,\xi_k)\|^2\right]+\mathbb{E}_{\xi_k}\left[\|\nabla F(q_k)-\nabla f(q_k,\xi_k)\|\right]\left(\|\nabla F(q_k)\|+\gamma\Delta\right)\Big)\\
&=: \alpha_k^2\lambda(I_1'+I_2').
\end{aligned}$$

The second-to-last equality holds because ξ_k and $\|\nabla F(q_k)\|$ are independent. By Assumption 4.ii), $I_1'\le\sigma^2$. Further, by Lemma B.1,

$$\|\nabla F(q_k)\| \le 2L_1(F(q_k)-F_*)+\frac{L_0}{L_1},$$

so that by Remark 3.3, we have

$$I_2' \le 2\sigma L_1(F(q_k)-F_*)+\frac{\sigma L_0}{L_1}+\sigma\gamma\Delta$$

and thus

$$\mathbb{E}_{\xi_k}[I_1] \le 2\alpha_k^2\lambda\sigma L_1(F(q_k)-F_*)+\alpha_k^2\lambda\sigma^2+\frac{\alpha_k^2\lambda\sigma L_0}{L_1}+\alpha_k^2\lambda\sigma\gamma\Delta. \tag{25}$$

We can bound the inner product in I_3 by noting that for any u,v it holds that

$$-\langle u,v\rangle = -\langle v,v\rangle+\langle v-u,v\rangle \le \|v\|^2+\|v-u\|\|v\|.$$

Thus,

$$\begin{aligned}
-\alpha_k\gamma\langle\nabla\varphi(p_{k+1}),\nabla\varphi(p_k)\rangle &\le -\alpha_k\gamma\|\nabla\varphi(p_k)\|^2+\alpha_k\gamma\|\nabla\varphi(p_{k+1})-\nabla\varphi(p_k)\|\|\nabla\varphi(p_k)\|\\
&\le \alpha_k^2\gamma\lambda\|\nabla f(q_k,\xi_k)-\gamma\nabla\varphi(p_k)\|\Delta.
\end{aligned}$$

By Remark 3.3, adding and subtracting $\nabla F(q_k)$, and applying Lemma B.1 we therefore get

$$\begin{aligned}
\mathbb{E}_{\xi_k}[I_3] &\le \frac{L_0}{2}\alpha_k^2\Delta^2+\frac{\alpha_k^2L_1\Delta^2}{2}\|\nabla F(q_k)\|+\alpha_k^2\gamma\lambda\Delta\left(\mathbb{E}_{\xi_k}\left[\|\nabla f(q_k,\xi_k)\|\right]+\gamma\|\nabla\varphi(p_k)\|\right)\\
&\le \frac{L_0}{2}\alpha_k^2\Delta^2+\Big(\frac{\alpha_k^2L_1\Delta^2}{2}+\alpha_k^2\gamma\lambda\Delta\Big)\|\nabla F(q_k)\|+\alpha_k^2\gamma\lambda\Delta\big(\mathbb{E}_{\xi_k}\left[\|\nabla f(q_k,\xi_k)-\nabla F(q_k)\|\right]+\gamma\Delta\big)\\
&\le \frac{L_0}{2}\alpha_k^2\Delta^2+\alpha_k^2\gamma\lambda\Delta(\sigma+\gamma\Delta)+\alpha_k^2\Delta\Big(\frac{L_1\Delta}{2}+\gamma\lambda\Big)\|\nabla F(q_k)\|\\
&\le \frac{L_0}{2}\alpha_k^2\Delta^2+\alpha_k^2\gamma\lambda\Delta(\sigma+\gamma\Delta)+\alpha_k^2\Delta^2\frac{L_0}{2}+\alpha_k^2\Delta\gamma\lambda\frac{L_0}{L_1}+\alpha_k^2\Delta L_1(L_1\Delta+2\gamma\lambda)(F(q_k)-F_*)
\end{aligned} \tag{26}$$

By gathering up the terms in the bounds of $\mathbb{E}_{\xi_k}[I_1]$ and $\mathbb{E}_{\xi_k}[I_3]$ from (25) and (26) we see that

$$\mathbb{E}_{\xi_k}[V_{k+1}]-V_k \le C_1(\alpha_k)+C_2(\alpha_k)(F(q_k)-F_*)$$

where

$$\begin{aligned}
C_1(\alpha_k) &= \alpha_k^2\lambda\sigma^2+\frac{\alpha_k^2\lambda\sigma L_0}{L_1}+\alpha_k^2\lambda\sigma\gamma\Delta+L_0\alpha_k^2\Delta^2+\alpha_k^2\gamma\lambda\Delta\sigma+\alpha_k^2\gamma^2\lambda\Delta^2+\alpha_k^2\Delta\gamma\lambda\frac{L_0}{L_1},\\
C_2(\alpha_k) &= 2\alpha_k^2\lambda\sigma L_1+\alpha_k^2\Delta^2L_1^2+2\alpha_k^2\gamma\lambda\Delta L_1.
\end{aligned}$$

Since $\varphi(p_k) - \varphi_* \geq 0$ it follows that $F(q_k) - F_* \leq H(p_k, q_k) - F_* - \varphi_*$, so in conclusion we have

$$\mathbb{E}_{\xi_k}\left[V_{k+1}\right] - V_k \leq C_1(\alpha_k) + C_2(\alpha_k)V_k, \tag{27}$$

Since $\sum_{k\geq 0} C_i(\alpha_k) < \infty$ for $i = 1, 2$, we can appeal to the Robbins–Siegmund theorem A.1 to conclude that $\lim_{k\to\infty} V_k$ exists and is finite almost surely. Since F and φ are coercive this implies that $\sup_k \|p_k\| < \infty$ and $\sup_k \|q_k\| < \infty$ almost surely.

□

Remark A.2. We note that the main reason for the stronger Assumption 4.ii) in Setting 2 compared to Assumption 3.ii) is to be able to bound the terms

$$\mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k, \xi_k)\|\right] \|\nabla F(q_k)\| \quad \text{and} \quad \mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k, \xi_k)\|^2\right]$$

which arise from the first scalar product of (23). In order to apply the Robbins-Siegmund theorem, these need to be bounded in terms of $F(q_k) - F_*$. As shown above, this is possible under Assumption 4.ii). But under Assumption 3.ii), these both lead to $\|\nabla F(q_k)\|^2$-terms. By Lemma B.1, we can therefore only guarantee a bound in terms of $(F(q_k) - F_*)^2$ in this setting. Unfortunately, Lemma B.1 can not be improved to bound $\|\nabla F(x)\|^2$ in terms of $F(q) - F_*$ (cf. [66, Lemma A.5]), this requires a stronger Lipschitz assumption like Assumption 3.i).

Under the weaker assumption that $\mathbb{V}\left[\nabla f(x, \xi)\right] \leq \kappa^2 \left(F(x) - F_*\right) + \tau^2\|\nabla F(x)\| + \sigma^2$, the second term $\mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k, \xi_k)\|^2\right]$ is bounded in terms of $F(q_k) - F_*$ and thus no longer an issue. However, while we can use the inequality $\sqrt{x} \leq x + 1$ to bound also $\mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k, \xi_k)\|\right]$ in terms of $F(q_k) - F_*$, this is not sufficient due to the multiplication by $\|\nabla F(q_k)\|$ which gives us another factor $F(q_k) - F_*$. Similarly, setting $\tau = 0$ and thus removing the $\nabla F(x)$-term altogether still leaves us with a bound involving $(F(q_k) - F_*)^{3/2}$. Removing the extra $(F(q_k) - F_*)^{1/2}$ requires us to also set $\kappa = 0$, which is equivalent to Assumption 4.ii).

Proof of Theorem 3.7 in Setting 3. Let F_* be as in Lemma B.2. By Lemma B.2, we have that the right-hand side of (24) can be bounded by

$$\alpha_k^2 \lambda \mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k, \xi_k)\|\right] \left(2L_1 N(F(q_k) - F_*) + \gamma\Delta\right).$$

By Assumption 5.ii) this can in its turn be bounded by

$$\alpha_k^2 \lambda\sigma \left(2L_1 N(F(q_k) - F_*) + \gamma\Delta\right).$$

The rest of the proof proceeds exactly like that of Setting 2 (with suitable modifications of the constants in the bound (27)). Similarly to the proof in Setting 1, we can also define

$$S_k = \frac{V_k}{\prod_{j=0}^{k-1}(1 + C_2(\alpha_j))}$$

and in a similar fashion obtain that $\sup_k \mathbb{E}\left[F(q_k)\right] < +\infty$. □

A.2. Integral form of (P_k, Q_k), part 1

To prove convergence, we start by considering the integral form of (9), namely

$$\begin{aligned}
p(t) &= p(0) - \int_0^t \nabla F(q(s))\mathrm{d}s - \gamma \int_0^t \nabla\varphi(p(s))\mathrm{d}s, \\
q(t) &= q(0) + \int_0^t \nabla\varphi(p(s))\mathrm{d}s.
\end{aligned} \tag{28}$$

Next, we rewrite the processes (13) on a form that is more reminiscent of this integral form. As in [31], we use the convention that

$$\sum_{i=n}^{k} a_i = 0, \quad \text{if } k = n-1 \text{ (the empty sum)},$$
$$\sum_{i=n}^{k} a_i = -\sum_{i=k+1}^{n-1} a_i, \quad \text{if } k < n-1.$$

By introducing the function

$$m(t) = \begin{cases} j, & t_j \le t < t_{j+1}, \\ 0, & t \le 0, \end{cases} \tag{29}$$

we can write (13) as

$$\begin{aligned} P_k(t) &= p_k + \sum_{i=k}^{m(t_k+t)-1} (p_{i+1} - p_i), \\ Q_k(t) &= q_k + \sum_{i=k}^{m(t_k+t)-1} (q_{i+1} - q_i). \end{aligned} \tag{30}$$

Using the fact that $p_k = P_k(0)$ and $q_k = Q_k(0)$, along with the update (10), we can rewrite (30) as

$$\begin{aligned} P_k(t) &= P_k(0) - \sum_{i=k}^{m(t_k+t)-1} \alpha_i \nabla F(q_i) + M_k(t) - \gamma \sum_{i=k}^{m(t_k+t)-1} \alpha_i \nabla \varphi(p_i), \\ Q_k(t) &= Q_k(0) + \sum_{i=k}^{m(t_k+t)-1} \alpha_i \nabla \varphi(p_{i+1}), \end{aligned} \tag{31}$$

where

$$M_k(t) = \sum_{i=k}^{m(t_k+t)-1} \alpha_i \delta M_i$$

and $\delta M_i = \nabla f(q_i, \xi_i) - \nabla F(q_i)$. In the next section, we show that the process $\{M_k\}_{k\ge 0}$ converges uniformly on compact sets, almost surely, to 0.

A.3. Convergence of the sequence $\{M_k\}$

The following lemma is an adaptation of part 1 of the proof of Theorem 2.1 from [31]:

Lemma A.3 (Convergence of $\{M_k(t)\}_{k\ge 0}$)**.** *Suppose that Assumption 1, 2 and 6 holds, along with either Assumption 3, 4 or 5. Then, the sequence $\{M_k(t)\}_{k\ge 0}$ converges uniformly on compact sets almost surely to 0. More precisely, for any T it holds that*

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \|M_k(t)\| = 0, \tag{32}$$

almost surely.

Proof. Closely following the proof of Theorem 2.1 in [31]: We let $\mathcal{F}_j = \sigma(\xi_1, \dots, \xi_j)$. By definition, we have that

$$M_k(t) = \sum_{i=k}^{m(t_k+t)-1} \alpha_i \delta M_i,$$

where $\delta M_i = \nabla f(q_i, \xi_i) - \nabla F(q_i)$. Define

$$\tilde{M}_j = \sum_{i=0}^{j-1} \alpha_i \delta M_i,$$

so that

$$M_k(t) = \tilde{M}_{m(t_k+t)} - \tilde{M}_k.$$

We will show that $\tilde{M}_j$ is a martingale sequence. We first note that

$$\mathbb{E}\left[\tilde{M}_{j+1} | \mathcal{F}_j\right] = \tilde{M}_j,$$

by Assumption 1.iii) and the fact that the noise is independent. Next, we demonstrate that

$$\mathbb{E}\left[\|\tilde{M}_j\|\right] < \infty, \tag{33}$$

Note that

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{M}_l\|^2\right] = \mathbb{E}\left[\left\|\sum_{i=0}^{l-1} \alpha_i \delta M_i\right\|^2\right] &= \mathbb{E}\left[\sum_{i=0}^{l-1} \alpha_i^2 \|\delta M_i\|^2 + 2\sum_{i=0}^{l-1}\sum_{j=0}^{i-1} \alpha_i \alpha_j \langle \delta M_i, \delta M_j\rangle\right] \\
&= \mathbb{E}\left[\sum_{i=0}^{l-1} \alpha_i^2 \|\delta M_i\|^2\right],
\end{aligned}$$

where we have used the fact that for $j < i$

$$\mathbb{E}\left[\langle \delta M_i, \delta M_j\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \delta M_i, \delta M_j\rangle | \mathcal{F}_j\right]\right] = \mathbb{E}\left[\langle \mathbb{E}\left[\delta M_i | \mathcal{F}_j\right], \delta M_j\rangle\right] = 0,$$

since M_j is $\mathcal{F}_j$-measurable and $\mathbb{E}\left[\delta M_i | \mathcal{F}_j\right] = 0$ (recall that ξ_i is independent of $\mathcal{F}_j$).

In the case that Assumption 4.ii) holds, we have that

$$\mathbb{E}\left[\|\delta M_i\|^2\right] < \infty. \tag{34}$$

Under Assumption 5, we get that (34) holds by Lemma B.4. If instead Assumption 3.ii) holds, we have that

$$\mathbb{E}\left[\|\delta M_{i+1} | \mathcal{F}_i\|^2\right] \leq \kappa(F(q_k) - F_*) + (1+\tau)\|\nabla F(q_k)\|^2 + \sigma^2.$$

By Assumption 3.i) and Lemma B.1, $\|\nabla F(q_k)\|^2 \leq 2L(F(q_k) - F_*)$, so by Theorem 3.7 the expectation of the right-hand side is finite. Thus (34) holds also in this case. Since this in particular means that also (33) is satisfied, we have shown that $\tilde{M}_j$ is a martingale in each of the considered settings.

We now show that (32) holds. By *Doob's submartingale inequality* [31, 60], we have for every $\mu > 0$ that

$$\mathbb{P}\left(\sup_{k \leq j \leq l} \|\tilde{M}_j - \tilde{M}_k\| \geq \mu\right) \leq \frac{\mathbb{E}\left[\|\tilde{M}_l - \tilde{M}_k\|^2\right]}{\mu^2} = \frac{\mathbb{E}\left[\|\sum_{i=k}^{l-1} \alpha_i \delta M_i\|^2\right]}{\mu^2},$$

which implies that

$$\mathbb{P}\left(\sup_{k \leq j} \|\tilde{M}_j - \tilde{M}_k\| \geq \mu\right) \leq C \sum_{i=k}^{\infty} \alpha_i^2,$$

and hence

$$\lim_{k\to\infty} \mathbb{P}\left(\sup_{k\leq j}\|\tilde{M}_j - \tilde{M}_k\| \geq \mu\right) = 0.$$

By Theorem 1(b) in [53, Section 2.10.3], the sequence $\{\tilde{M}_j\}_{j\geq 0}$ is thus a Cauchy sequence almost surely, and by Theorem 4 in [53, Section 2.10.5] therefore converges almost surely. Thus $\lim_{k\to\infty} \sup_{k\leq j}\|\tilde{M}_j - \tilde{M}_k\| = 0$ holds almost surely. Since for any interval $[0,T]$ we have that

$$\sup_{t\in[0,T]} \|M_k(t)\| = \sup_{t\in[0,T]} \|\tilde{M}_{m(t_k+t)} - \tilde{M}_k\| = \sup_{k\leq j\leq m(t_k+t)} \|\tilde{M}_j - \tilde{M}_k\| \leq \sup_{k\leq j}\|\tilde{M}_j - \tilde{M}_k\|,$$

the lemma follows. □

A.4. Equicontinuity of the sequences $\{P_k\}_{k\geq 0}$ and $\{Q_k\}_{k\geq 0}$

The goal of this section is to show the following lemma.

Lemma 3.9 (Equicontinuous in the extended sense)**.** *Consider $\{Z_k\}_{k\geq 0} = \{(P_k, Q_k)\}_{k\geq 0}$ where the sequences $\{P_k\}_{k\geq 0}$ and $\{Q_k\}_{k\geq 0}$ are defined by (13) (equivalently, by (30)). Suppose that $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$ are defined by (10), and that the Hamiltonian is on the form (7). Further, let Assumptions 1, 2 and 6 be valid, as well as either Assumption 3, 4 or 5. Then $\{Z_k\}_{k\geq 0}$ is equicontinuous in the extended sense, almost surely.*

This will be accomplished by instead verifying the following equivalent condition:

Lemma A.4 (Equivalent definition of extended equicontinuity)**.** *A sequence of functions $\{f_k\}_{k\geq 0}$, $f_k : \mathbb{R} \to \mathbb{R}^d$ is* equicontinuous in the extended sense *if and only if $\{\|f_k(0)\|\}_{k\geq 0}$ is bounded and for every T and $\epsilon > 0$ there exists a $\delta > 0$ and a null sequence $(a_k)_{k\geq 0}$ (that is, $\lim_{k\to\infty} a_k = 0$) such that*

$$\sup_{\substack{0<|t-s|\leq\delta,\\ t,s\in[0,T]}} \|f_k(t) - f_k(s)\| \leq \epsilon + a_k. \tag{35}$$

Proof of Lemma A.4. By definition, (14) is equal to

$$\lim_{k\to\infty} b_k \leq \epsilon$$

with

$$b_k := \sup_{j\geq k} \sup_{\substack{0<|t-s|\leq\delta,\\ t,s\in[0,T]}} \|f_j(t) - f_j(s)\|.$$

Define

$$a_k = \max\{0, b_k - \epsilon\}.$$

Then $(a_k)_{k\geq 0}$ satisfies all the requirements; a_k is clearly positive and by continuity of the function $\max\{0,x\}$ it holds that

$$\lim_{k\to\infty} a_k = \max\{0, \lim_{k\to\infty} b_k - \epsilon\} = 0,$$

as $\lim_{k\to\infty} b_k - \epsilon \leq 0$. Furthermore, we have that

$$\sup_{\substack{0<|t-s|\leq\delta,\\ t,s\in[0,T]}} \|f_k(t) - f_k(s)\| \leq b_k \leq \epsilon + a_k,$$

for every k and thus we have shown that (35) follows from (14). We now show the converse. Suppose that (35) holds. Taking the supremum of (35) we obtain

$$\sup_{j\geq k}\sup_{\substack{0<|t-s|\leq\delta,\\ t,s\in[0,T]}}\|f_j(t)-f_j(s)\|\leq\epsilon+\sup_{j\geq k}a_j.$$

We finally take the limit with respect to k

$$\lim_{k\to\infty}\sup_{j\geq k}\sup_{\substack{0<|t-s|\leq\delta,\\ t,s\in[0,T]}}\|f_j(t)-f_j(s)\|\leq\epsilon+\lim_{k\to\infty}\sup_{j\geq k}a_j=:\epsilon+\limsup_{k\to\infty}a_k.$$

But as $\lim_{k\to\infty}a_k$ exists by assumption and equals 0, $\limsup_{k\to\infty}a_k=\lim_{k\to\infty}a_k=0$. We thus conclude that (14) holds. $\square$

We now turn to the proof of Lemma 3.9.

Proof of Lemma 3.9. Closely following Lemma 2 in [20]: We want to show that the sequence $\{Z_k\}_{k\geq0}=\{(P_k,Q_k)\}_{k\geq0}$, where $\{P_k\}_{k\geq0}$ and $\{Q_k\}_{k\geq0}$ are defined by (13), is equicontinuous in the extended sense.

First, we note that the sequences $\{P_k(0)\}$ and $\{Q_k(0)\}$ are finite except on a set of measure 0, since by Theorem 3.7 $\sup_k\|p_k\|<\infty$ and $\sup_k\|q_k\|<\infty$ almost surely.

By Lemma A.4 we are done if for every $\epsilon>0$, there is a sequence $\{a_k\}_{k\geq0}$ such that $\lim_{k\to\infty}a_k=0$ and a $\delta>0$ such that

$$\sup_{\substack{|t-s|<\delta,\\ t,s\in[0,T]}}\|Z_k(t)-Z_k(s)\|\leq\epsilon+a_k,\ \text{a.s.} \tag{36}$$

By (31), we have that

$$\|P_k(t)-P_k(s)\|\leq C(\omega)\sum_{i=m(t_k+s)}^{m(t_k+t)-1}\alpha_i+\|M_k(t)\|+\|M_k(s)\|, \tag{37}$$

where $C(\omega)=\sup_i\|\nabla F(q_i)-\gamma\nabla\varphi(p_i)\|$. By the boundedness of p_k and q_k along with the continuity of ∇F and $\nabla\varphi$, we have that $C(\omega)<\infty$, a.s. The sum on the right-hand side of (37) can be rewritten as

$$\sum_{i=m(t_k+s)}^{m(t_k+t)-1}\alpha_i=t_{m(t_k+t)}-t_{m(t_k+s)}.$$

By definition of m, (29), we have that

$$t_{m(t_k+t)}-t_{m(t_k+s)}\leq t_k+t-t_{m(t_k+s)}. \tag{38}$$

But we also have

$$t_{m(t_k+s)}\leq t_k+s<t_{m(t_k+s)+1},$$

and hence the right-hand side of (38) can be rewritten and bounded as follows

$$t_k+t-(t_k+s)+(t_k+s)-t_{m(t_k+s)}\leq(t-s)+t_{m(t_k+s)+1}-t_{m(t_k+s)}.$$

Now, $t_{m(t_k+s)+1}-t_{m(t_k+s)}=\alpha_{m(t_k+s)+1}$ and hence we see that

$$\|P_k(t)-P_k(s)\|\leq C(\omega)\left(|t-s|+\alpha_{m(t_k+s)+1}\right)+\|M_k(t)\|+\|M_k(s)\|.$$

Let ϵ be greater than 0. There are now two cases. If $C(\omega) = 0$, (35) clearly holds for any $\delta > 0$. If $C(\omega) \neq 0$, then take $\delta > 0$ so small that $C(\omega)\delta < \epsilon$. We then have

$$\begin{aligned}\sup_{\substack{|t-s|<\delta,\\ t,s\in[0,T]}} \|P_k(t) - P_k(s)\| &\leq \sup_{\substack{|t-s|<\delta,\\ t,s\in[0,T]}} \big(C(\omega)(|t-s| + \alpha_{m(t_k+s)+1}) + \|M_k(t)\| + \|M_k(s)\|\big) \\ &< \epsilon + C(\omega)\alpha_{m(t_k)+1} + 2\|M_k(T)\|.\end{aligned}$$

By Lemma A.3, $\lim_{k\to\infty}\|M_k(T)\| = 0$ a.s. and we see that (35) in Lemma A.4 holds almost surely. A similar argument yields an analogous bound for Q_k, and by the equivalence of norms on $\mathbb{R}^{2d}$, we obtain (36). $\square$

A.5. Integral form of (P_k, Q_k), part 2

We now use the extended equicontinuity established in Lemma 3.9 to continue the reformulation of (P_k, Q_k) into a form closer to (28). The following lemma extends (31):

Lemma A.5 (Integral form of (P_k, Q_k))**.** *Under the same assumptions and notation as in Lemma 3.9, we can write*

$$\begin{aligned}P_k(t) &= P_k(0) - \int_0^t \nabla F(Q_k(s))\mathrm{d}s - \gamma\int_0^t \nabla\varphi(P_k(s))\mathrm{d}s + M_k(t) + \mu_k(t),\\ Q_k(t) &= Q_k(0) + \int_0^t \nabla\varphi(P_k(s))\mathrm{d}s + \nu_k(t) + \kappa_k(t),\end{aligned} \tag{39}$$

where the functions $\{M_k\}_{k\geq0}$, $\{\mu_k\}_{k\geq0}$, $\{\nu_k\}_{k\geq0}$ and $\{\kappa_k\}_{k\geq0}$ converge to 0 uniformly on compact sets almost surely.

Proof. We start by showing that the first sum in Equation (31) satisfies

$$-\sum_{i=k}^{m(t_k+t)-1} \alpha_i \nabla F(q_i) = -\int_0^t \nabla F(Q_k(s))\mathrm{d}s + \mu_{1,k}(t),$$

where $\{\mu_{1,k}\}_{k\geq0}$ is a sequence of functions that tends to 0 uniformly on compact intervals. Consider

$$I_k := -\int_0^t \nabla F(Q_k(s))\mathrm{d}s.$$

Then, since $t_k + s$ belongs to a single interval $[t_i, t_{i+1})$,

$$\begin{aligned}I_k &= -\int_0^t \left(\nabla F(q_0) I_{(-\infty,t_0)}(t_k+s) - \sum_{i=0}^{\infty} \nabla F(q_i) I_{[t_i,t_{i+1})}(t_k+s)\right)\mathrm{d}s\\ &= -\int_0^t \left(\nabla F(q_0) I_{(-\infty,t_0-t_k)}(s) - \sum_{i=0}^{\infty} \nabla F(q_i) I_{[t_i-t_k,t_{i+1}-t_k)}(s)\right)\mathrm{d}s.\end{aligned}$$

Since $s \in [0, t]$ and $t_0 - t_k \leq 0$, the first term disappears. For $i < k$, we have that $t_{i+1} - t_k \leq 0$. We can therefore start the sum at $i = k$, as earlier terms will not contribute to the integral. Thus,

$$I_k = -\int_0^t \left(\sum_{i=k}^{\infty} \nabla F(q_i) I_{[t_i-t_k,t_{i+1}-t_k)}(s)\right)\mathrm{d}s.$$

Now suppose that $t_j - t_k \leq t < t_{j+1} - t_k$. We split up the previous integral as follows:

$$\begin{aligned} I_k &= -\int_0^{t_j - t_k} \left(\sum_{i=k}^{j-1} \nabla F(q_i) I_{[t_i - t_k, t_{i+1}-t_k)}(s) \right) \mathrm{d}s - \int_{t_j - t_k}^{t} \nabla F(q_k) I_{[t_j - t_k, t_{j+1} - t_k)}(s) \mathrm{d}s \\ &= -\sum_{i=k}^{j-1} \nabla F(q_i)\alpha_i - \nabla F(q_j)\left(t - t_j + t_k\right), \end{aligned}$$

where we have used that $\int_0^{t_j - t_k} I_{[t_i - t_k, t_{i+1} - t_k)}(s)\mathrm{d}s = \alpha_i$. Using the fact that $m(t + t_k) = j$, we can rewrite this further as

$$I_k = -\sum_{i=k}^{m(t_k+t)-1} \nabla F(q_i)\alpha_i - \mu_{1,k}(t),$$

where $\mu_{1,k}(t) = \nabla F(q_{m(t_k+t)})\left(t - t_{m(t_k+t)} + t_k\right)$. The function $\mu_{1,k}$ is piecewise linear and 0 at $t = t_j - t_k$. Since $\sup_k \|q_k\| < \infty$ a.s. by Theorem 3.7, the set $\{q_k\}$ can be enclosed in a compact set. As ∇F is continuous in each of the settings that we consider, a positive random variable $C(\omega)$ which satisfies

$$\|\nabla F(q_{m(t_k+t)})\| \leq C(\omega) < \infty, \text{ a.s.},$$

thus exists. Hence, it holds that

$$\|\mu_{1,k}(t)\| \leq C(\omega)|\alpha_{m(t_k+t)}|.$$

Now, we have that $\lim_{k\to\infty} \sup_{t\in[0,T]} \alpha_{m(t_k+t)} = 0$ for any fixed T, since $\lim_{k\to\infty}\alpha_k = 0$, and thus $\mu_{1,k}$ converges to 0 uniformly on compact intervals.

In a similar fashion we obtain for the last term in (31) that

$$-\sum_{i=k}^{m(t_k+t)-1} \alpha_i \nabla\varphi(p_i) = -\int_0^t \nabla\varphi(P_k(s))\mathrm{d}s + \mu_{2,k}(t),$$

where $\{\mu_{2,k}\}_{k\geq 0}$ converges uniformly on compact sets to 0. Letting $\mu_k = \mu_{1,k} + \gamma\mu_{2,k}$, we obtain the expression in the first line of (39).

We now turn our attention to the second line of (39). By an argument analogous to the previous, we can write the second line of (31) as

$$Q_k(t) = Q_k(0) + \int_0^t \varphi(P_{k+1}(s))\mathrm{d}s + \nu_k(t),$$

where ν_k converges uniformly on compact sets to 0. We can rewrite the integral on the right-hand side as

$$\int_0^t \nabla\varphi(P_{k+1}(s))\mathrm{d}s = \underbrace{\int_0^t \nabla\varphi(P_{k+1}(s)) - \nabla\varphi(P_k(s))\mathrm{d}s}_{:=\kappa_k(t)} + \int_0^t \nabla\varphi(P_k(s))\mathrm{d}s.$$

By the Lipschitz continuity of $\nabla\varphi$ and the definition of P_k (13),

$$\|\kappa_k(t)\| \leq \int_0^t \lambda\|P_{k+1}(s) - P_k(s)\|\mathrm{d}s = \int_0^t \lambda\|P_k(\alpha_k + s) - P_k(s)\|\mathrm{d}s.$$

Since $\{P_k(t)\}_{k\geq 0}$ is equicontinuous in the extended sense by Lemma 3.9, there is for each T and $\epsilon > 0$ a $\delta > 0$ such that

$$\limsup_{k\to\infty} \sup_{|t-s|<\delta,\ t,s\in[0,T]} \|P_k(\alpha_k + s) - P_k(s)\| \leq \epsilon. \tag{40}$$

What remains is to show that $\{\kappa_k\}_{k\geq 0}$ converges uniformly on compact sets to 0. For any T, we have that

$$\lim_{k\to\infty}\sup_{t\in[0,T]}\|\kappa_k(t)\| \leq \lim_{k\to\infty}\int_0^T \lambda\|P_k(\alpha_{k+1}+s)-P_k(s)\|\mathrm{d}s$$

since the integrand is positive. By Theorem 3.7, we can bound $\|P_k(\alpha_{k+1}+s)-P_k(s)\| \leq 2\sup_{t\in\mathbb{R}}\|P_k(t)\| < \infty$. Thus, we can use the Lebesgue dominated convergence theorem and take the limit inside the integral:

$$\lim_{k\to\infty}\sup_{t\in[0,T]}\|\kappa_k(t)\| \leq \int_0^T \lim_{k\to\infty}\lambda\|P_k(\alpha_{k+1}+s)-P_k(s)\|\mathrm{d}s.$$

By (40), we can make the integrand arbitrarily small by choosing k so large that $\alpha_{k+1}\leq\delta$. □

In the next section, we use the integral form of the processes $\{P_k\}_{k\geq 0}$ and $\{Q_k\}_{k\geq 0}$ to establish that (P_0, Q_0) is an asymptotic pseudotrajectory for (9). This essentially means that (P_k, Q_k) asymptotically satisfies (28).

A.6. **Asymptotic pseudotrajectory**

Definition A.6. A function $X:\mathbb{R}\to\mathbb{R}^{2d}$ is an asymptotic pseudotrajectory to a semiflow ϕ if

$$\lim_{s\to\infty}\sup_{t\in[0,T]}\|X(s+t)-\phi(t,X(s))\| = 0,$$

for all $T>0$.

Remark A.7. Definition A.6 differs from the definition encountered in e.g. [7,8] in that it requires the function to be continuous. In our case, the extended equicontinuity (Definition 3.8 plays a similar role (see the proof of Lemma 3.10, but we omit this from the definition.

Lemma 3.10 (Asymptotic pseudotrajectory)**.** *Let Assumptions 1, 2 and 6 be valid, as well as either Assumption 3, 4 or 5. Then the function $Z_0 = (P_0, Q_0)$ where P_0 and Q_0 are defined by (12) is almost surely an asymptotic pseudotrajectory to the semiflow ϕ generated by equation (9). That is,*

$$\lim_{s\to\infty}\sup_{t\in[0,T]}\|Z_0(s+t)-\phi(t,Z_0(s))\| = 0,$$

for all $T>0$.

Proof. It is sufficient to show the lemma under the (L_0, L_1)-smoothness assumption given by Assumption 4.i), since Assumptions 3.i) and 5.iii) are special cases of this. We have by (39) that

$$\begin{aligned}\|Z_k(t)-\phi(t,z_k)\| &\leq \|P_k(t)-\phi_p(t,p_k)\| + \|Q_k(t)-\phi_q(t,q_k)\| \\ &= \left\|\int_0^t \nabla F\left(Q_k(s)-\phi_q(s,p_k)\right)\mathrm{d}s\right\| + \left\|\int_0^t \nabla\varphi\left(P_k(s)-\phi_p(s,p_k)\right)\mathrm{d}s + M_k(t)+\mu_k(t)\right\| \\ &\quad + \left\|\int_0^t \nabla\varphi\left(P_k(s)-\phi_p(s,p_k)\right)\mathrm{d}s + \nu_k(t)+\kappa_k(t)\right\|\end{aligned} \tag{41}$$

By Assumption 4.i), the first term satisfies

$$\left\|\int_0^t \nabla F\left(Q_k(s)-\phi_q(s,p_k)\right)\mathrm{d}s\right\| \leq \int_0^t \left(L_0 + L_1\|\nabla F\left(Q_k(s)\right)\|\right)\|Q_k(s)-\phi_q(s,p_k)\|\mathrm{d}s.$$

For the $\|\nabla F(Q_k(s))\|$-term, we note that this is less than or equal to $\sup_{k\geq 0}\|\nabla F(q_k)\|$, which is bounded almost surely by Theorem 3.7 and the continuity of ∇F. Using the Lipschitz continuity of $\nabla\varphi$ we get that the right-hand side of (41) is less than or equal to

$$\int_0^t \left(L_0 + L_1 \sup_{k\geq 0}\|\nabla F(q_k)\|\right)\|Q_k(s) - \phi_q(s, q_k)\|\mathrm{d}s + \int_0^t 2\lambda\|P_k(s) - \phi_p(s, p_k)\|\mathrm{d}s \\ + \|M_k(t)\| + \|\mu_k(t)\| + \|\nu_k(t)\| + \|\kappa_k(t)\|.$$

Hence, we can find a constant C such that

$$\|Z_k(t) - \phi(t, z_k)\| \leq \int_0^t C\|Z_k(s) - \phi(s, z_k)\|\mathrm{d}s + \|M_k(t)\| + \|\mu_k(t)\| + \|\nu_k(t)\| + \|\kappa_k(t)\|.$$

By Grönwall's inequality, compare [15], it holds that

$$\|Z_k(t) - \phi(t, z_k)\| \leq (\|M_k(t)\| + \|\mu_k(t)\| + \|\nu_k(t)\| + \|\kappa_k(t)\|)\, e^{Ct}.$$

Since all the terms within the brackets on the right-hand side converge to 0 uniformly on compact intervals, it follows that

$$\lim_{k\to\infty}\sup_{t\in[0,T]}\|Z_k(t) - \phi(t, z_k)\| = 0,$$

almost surely. Thus the second term in the right-hand-side of the inequality

$$\begin{aligned}\sup_{t\in[0,T]}\|Z_0(s+t) - \phi(t, Z_0(s))\| \leq &\sup_{t\in[0,T]}\|Z_0(s+t) - Z_0(t_{m(s)}+t)\| \\ &+ \sup_{t\in[0,T]}\|Z_0(t_{m(s)}+t) - \phi(t, Z_0(t_{m(s)}))\| \\ &+ \sup_{t\in[0,T]}\|\phi(t, Z_0(t_{m(s)})) - \phi(t, Z_0(s))\|\end{aligned}$$

tends to 0. The first term also tends to 0 as $s\to\infty$ since Z_0 is continuous in the extended sense. Finally, the third term tends to 0 by the continuity of ϕ. It follows that Z_0 satisfies

$$\lim_{s\to\infty}\sup_{t\in[0,T]}\|Z_0(s+t) - \phi(t, Z_0(s))\| = 0,$$

for all $T > 0$, almost surely, and thus Z_0 is an asymptotic pseudotrajectory to ϕ. □

A.7. Chain transitivity

Definition A.8 (Chain transitivity)**.** An (ϵ, T)-*pseudo orbit* from x to y is a sequence of points $\{x_i\}_{i=0}^n$ and time points $t_i \geq T$, such that $x_0 = x$ and $x_n = y$ and

$$\|\phi(t_i, x_i) - x_{i+1}\| < \epsilon.$$

An invariant set A is said to be *internally chain transitive* if for each $a, b \in A$ and $T > 0, \epsilon > 0$ there exists an (ϵ, T)-pseudo orbit from a to b.

The following result is essentially a paraphrasing of [7, Corollary 4.3] combined with [8, Proposition 5.3]:

Lemma 3.11 (Internally chain transitive)**.** *Let Assumptions 1, 2 and 6 be valid, as well as either Assumption 3, 4 or 5. Then given that* $\{z_k\}_{k\geq 0} = \{(p_k, q_k)\}_{k\geq 0}$ *is determined by* (10)*, the set* $\mathrm{Lim}(\{z_k\}_{k\geq 0})$ *is a compact, and internally chain transitive set for the semiflow* ϕ *generated by* (9)*.*

Proof of Lemma 3.11. The proof is very similar to the proof of Corollary 4.3 in [7], except that our Z_0 is piece-wise constant rather than the linear interpolation in [7]. This influences how $\mathrm{Lim}(Z_0) = \mathrm{Lim}(\{z_k\}_{k\geq 0})$ is established and the argument for why $\mathrm{Lim}(\{z_k\}_{k\geq 0})$ is connected and compact. But the former follows immediately from the fact that Z_0 only takes on the values of $\{z_k\}_{k\geq 0}$, and by [3], the set of limit points is connected for almost every ω. Since it is also closed and bounded by Theorem 3.7, it is compact for almost all ω. The rest of the proof proceeds like in [7], making use of Lemma 3.10. This shows that $\mathrm{Lim}(\{z_k\}_{k\geq 0})$ is a connected *internally chain recurrent* set. But by [8, Proposition 5.3], such a set is internally chain transitive. $\square$

A.8. **Hamiltonian dynamics**

By Lemma 3.11, $\mathrm{Lim}(\{z_k\}_{k\geq 0})$ is internally chain transitive. We will now relate $\mathrm{Lim}(\{z_k\}_{k\geq 0})$ to E_H by using that we have a Lyapunov function for the flow generated by (9). Let us first specify exactly what we mean by a Lyapunov function:

Definition A.9 (Lyapunov function)**.** Let ϕ be a semiflow on a metric space W and let $\Lambda \subset W$ be an invariant set for ϕ. A function $V : W \to \mathbb{R}$ is called a *Lyapunov function* for Λ if the function $t \in \mathbb{R}_+ \mapsto V(\phi(t, x))$ is constant for $x \in \Lambda$ and eventually decreasing for $x \notin \Lambda$.

Remark A.10. In [8], the above definition is stated as "strictly decreasing" rather than "eventually decreasing". However, the results in [8] which we will use only require the latter. The critical part is that the Lyapunov function does not remain constant for all time when $x \notin \Lambda$.

Lemma 3.12 (Lyapunov function)**.** *Let Assumptions 1 and 2 be valid, and let $V : \mathbb{R}^{2d} \to \mathbb{R}$ be defined by*

$$V(z) = H(p, q) - F_* - \varphi_*, \tag{15}$$

where $z = (p, q)$. Then V is a Lyapunov function for $E_H \subset \mathbb{R}^{2d}$ with respect to ϕ.

Proof. Let $z(t) = (p(t), q(t))$ denote the solution to (9) with initial condition $z_0 = (p_0, q_0)$. The function V is constant if and only if $H(p(t), q(t))$ is constant, and we have

$$\frac{\mathrm{d}}{\mathrm{d}t} H(p(t), q(t)) = -\gamma \|\nabla\varphi(p(t))\|^2.$$

That $z_0 \in E_H$ means that $\nabla F(q_0) = 0$ and $\nabla\varphi(p_0) = 0$. Then by (9) also $\dot{z}(t) = 0$, and in particular $\nabla\varphi(p(t)) = 0$. Thus V is constant on E_H. Outside E_H there are two cases; either $\nabla\varphi(p_0) \neq 0$ or $\nabla F(q_0) \neq 0$. In the former case, $\frac{\mathrm{d}}{\mathrm{d}t} H(p(t), q(t))|_{t=0} < 0$, so V is decreasing. In the latter case, we could potentially have the situation that $\nabla\varphi(p(t)) = 0$ for all $t \geq 0$ (and thus $q(t) \equiv q_0$) but $\nabla F(q_0) \neq 0$, so that $(p(t), q_0) \notin E_H$ for any t. However, this is precluded by the convexity and coercivity assumptions on φ, in the following way. Since $\nabla\varphi(p(t)) = 0$ and $q(t) = q_0$ is constant, we get $\dot{p}(t) = -\nabla F(q_0)$, so that

$$p(t) = p_0 - \nabla F(q_0)t.$$

Thus $\|p(t)\| \to \infty$ as $t \to \infty$, which means that also $\|\varphi(p(t))\| \to \infty$ as $t \to \infty$ since φ is coercive. But by the convexity of $\nabla\varphi$ we have that for any $x \in \mathbb{R}^n$,

$$\varphi(x) \geq \varphi(p(t)) + \langle \nabla\varphi(p(t)), x - p(t)\rangle = \varphi(p(t)),$$

which implies that $\varphi(x) = \infty$ for every $x \in \mathbb{R}^n$. This contradicts Assumption 2, and thus there must exist a $t > 0$ such that $\nabla\varphi(t) \neq 0$. This means that V is eventually decreasing.

$\square$

A.9. **Proof of the main theorem**

We will now make use of the following proposition from [8] to prove the final result:

Proposition A.11 (Proposition 6.4 in [8])**.** *Let ϕ be a semiflow on a metric space W and $\Lambda \subset W$ a compact, invariant set. Let $V : W \to \mathbb{R}$ be a Lyapunov function for Λ and suppose that $V(\Lambda)$ has empty interior. Then every internally chain transitive set is contained in Λ.*

Theorem 3.5. *Let Assumptions 1, 2 and 6 be satisfied, as well as either Assumption 3, 4 or 5. Then $\{q_k\}_{k\geq 0}$ converges almost surely to the set of stationary points of the objective function F. If we additionally assume that Assumption 7 holds, the convergence is to a unique stationary point.*

Proof. Let ϕ be the semiflow generated by (9). The following statements all hold almost surely. By Lemma 3.11, $\mathrm{Lim}(\{z_k\}_{k\geq 0})$ is a compact and internally chain transitive set for ϕ. It is therefore in particular an invariant set for ϕ. Now set $\Lambda = E_H \cap \mathrm{Lim}(\{z_k\}_{k\geq 0})$. Then Λ is also invariant for ϕ, since E_H clearly is and the intersection of two invariant sets is invariant. Further since E_H is closed and $\mathrm{Lim}(\{z_k\}_{k\geq 0})$ is compact, Λ is compact. Finally, Λ is non-empty, because for any $z \in \mathrm{Lim}(\{z_k\}_{k\geq 0})$, $\{\phi(t,z)\}_{t\geq 0}$ is a compact subset of $\mathrm{Lim}(\{z_k\}_{k\geq 0})$. We can thus extract a subsequence such that $\mathrm{Lim}(\phi) \ni \bar{z} = \lim_{n\to\infty} \phi(t_n, z) \in \mathrm{Lim}(\{z_k\}_{k\geq 0})$. Since $\mathrm{Lim}(\phi) \subset E_H$ by the same argument as in [59, Theorem 15.0.3], $\bar{z} \in \Lambda \neq \emptyset$.

Due to the invariance of $\mathrm{Lim}(\{z_k\}_{k\geq 0})$, we can consider the restricted semiflow $\Psi = \phi|_{\mathrm{Lim}(\{z_k\}_{k\geq 0})}$, and it follows from Proposition 3.12 that the function V defined in (15) is a Lyapunov function for Λ with respect to Ψ. Since $\mathrm{Lim}(\{z_k\}_{k\geq 0})$ is internally chain transitive by Lemma 3.11, and $V(\Lambda)$ has empty interior by Assumption 1.ii). Proposition A.11 then shows that

$$\mathrm{Lim}(\{z_k\}_{k\geq 0}) \subset \Lambda \subset E_H.$$

This proves the main assertion of Theorem 3.5. Under Assumption 7, the set E_H consists of isolated points and clearly the convergent sequence can only approach one of these points arbitrarily closely. □

Corollary 3.6 (Convergence in expectation)**.** *Let Assumptions 1, 2, 6 and 7 be valid. Further, let the Hamiltonian be on the form (7) and let the sequences $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$ be generated by (10). Then it holds that*

$$\lim_{k\to\infty} \mathbb{E}\left[\|\nabla F(q_k)\|^{\theta}\right] = 0,$$

where $0 < \theta < 2$ under Assumption 3 and $0 < \theta < 1$ under Assumption 4 or 5.

Proof. From Lemma B.1 we have that $\|\nabla F(q_k)\|^2 \leq 2L(F(q_k) - F_*)$ under Assumption 3. By Theorem 3.7 we therefore obtain that

$$\sup_k \mathbb{E}\left[\|\nabla F(q_k)\|^2\right] < \infty.$$

By Lemma 3 in [53, Chapter 2.6] with $G(t) = t^{2/\theta}$ and $\xi_n = \|\nabla F(q_n)\|^{\theta}$, the sequence $\{\|\nabla F(q_k)\|^{\theta}\}_{k\geq 0}$ is uniformly integrable for $0 < \theta < 2$. Since $q_k \to q_*$ a.s. by Theorem 3.5 and $\nabla F(q_*) = 0$ we also have that $\|\nabla F(q_k)\|^{\theta} \to 0$ a.s. Thus by Theorem 4 in [53, Chapter 2.6], $\mathbb{E}\left[\|\nabla F(q_k)\|^{\theta}\right] \to 0$. The argument is the same under Assumption 4 or 5, but in these cases Lemma B.1 only guarantees that

$$\sup_k \mathbb{E}\left[\|\nabla F(q_k)\|\right] < \infty.$$

The sequence $\{\|\nabla F(q_k)\|^{\theta}\}_{k\geq 0}$ is therefore only guaranteed to be uniformly integrable when $0 < \theta < 1$. □

Appendix B. Auxiliary results

The following results summarize important consequences of the assumptions in Settings 1–3.

Lemma B.1. *Let F be bounded from below by F_*. If F is (L_0, L_1)-smooth, i.e. satisfies Assumption 4.i), it holds that*

$$\|\nabla F(q)\| \leq 2L_1(F(q) - F_*) + \frac{L_0}{L_1}.$$

If F is instead L-smooth with Lipschitz constant L, i.e. satisfies Assumption 3.i), it holds that

$$\|\nabla F(q)\|^2 \leq 2L(F(q) - F_*).$$

Proof. Consider first the (L_0, L_1)-smooth case. Put

$$q_+ = q - \frac{1}{L_1\|\nabla F(q)\|}\nabla F(q), \tag{42}$$

so that $\|q_+ - q\| = \frac{1}{L_1}$ and the conditions for (21) in [66] are satisfied. Thus it holds that

$$F(q_+) - F(q) \leq \langle \nabla F(q), q_+ - q\rangle + \frac{L_0 + L_1\|\nabla F(q)\|}{2}\|q_+ - q\|^2.$$

Inserting (42) into the previous expression we see that

$$F(q_+) - F(q) \leq -\frac{1}{L_1}\|\nabla F(q)\| + \frac{L_0 + L_1\|\nabla F(q)\|}{2}\frac{1}{L_1^2},$$

Rearranging the terms, we find that

$$F(q_+) - F(q) \leq -\frac{1}{2L_1}\|\nabla F(q)\| + \frac{L_0}{2}\frac{1}{L_1^2}$$

One more rearrangement yields

$$\|\nabla F(q)\| \leq 2L_1(F(q) - F(q_+)) + \frac{L_0}{L_1}.$$

Since $F(q_+) \geq F_*$ we obtain the statement of the first part of the Lemma. The proof of the second part is very similar but simpler, and therefore omitted. □

Lemma B.2. *Let F satisfy Assumption 5.i) and $f(\cdot, \xi)$ satisfy Assumption 5.iii). Then*

$$\|\nabla f(q, \xi)\| \leq 2NL_1\left(F(q) - F_*\right) + \frac{L_0}{L_1}, \tag{43}$$

almost surely, where $F_* = \frac{1}{N}\sum_{i=1}^{N}\inf_{q\in\mathbb{R}^d} f_i(x)$.

Proof. We start by showing that there exists a constant C such that

$$f(x, \xi) - \inf_{x\in\mathbb{R}^d} f(x, \xi) \leq C(F(x) - F_*), \tag{44}$$

almost surely. First we note that by the properties of inf it holds that

$$\inf_{x\in\mathbb{R}^d} f(x, \xi) = \inf_{x\in\mathbb{R}^d}\left(\frac{1}{|B_\xi|}\sum_{i\in B_\xi} f_i(x)\right) = \frac{1}{|B_\xi|}\cdot\inf_{x\in\mathbb{R}^d}\left(\sum_{i\in B_\xi} f_i(x)\right) \geq \frac{1}{|B_\xi|}\cdot\sum_{i\in B_\xi}\inf_{x\in\mathbb{R}^d} f_i(x).$$

Hence

$$f(x,\xi) - \inf_{x\in\mathbb{R}^d} f(x,\xi) = \frac{1}{|B_\xi|}\cdot\sum_{i\in B_\xi} f_i(x) - \inf_{x\in\mathbb{R}^d} f(x,\xi)$$
$$\leq \frac{1}{|B_\xi|}\cdot\sum_{i\in B_\xi} f_i(x) - \frac{1}{|B_\xi|}\cdot\sum_{i\in B_\xi}\inf_{x\in\mathbb{R}^d} f_i(x).$$

Since $f_i(x) - \inf_{x\in\mathbb{R}^d} f_i(x)) \geq 0$, the previous expression can be bounded by

$$\frac{1}{|B_\xi|}\cdot\sum_{i=1}^{N} f_i(x) - \inf_{x\in\mathbb{R}^d} f_i(x) = \frac{N}{|B_\xi|}\left(F(x) - F_*\right).$$

As $1 \leq |B_\xi| \leq N$, (44) holds. Since $f(\cdot,\xi)$ is (L_0, L_1)-smooth, it further holds that

$$\|\nabla f(x,\xi)\| \leq 2L_1\left(f(x,\xi) - \inf_{x\in\mathbb{R}^d} f(x,\xi)\right) + \frac{L_0}{L_1}.$$

Combining the previous expression with (44), we obtain (43). □

Remark B.3. Note that for fixed, *deterministic* x, (43) implies that the norm $\|\nabla f(x,\xi) - \nabla F(x)\|$ is bounded almost surely. This means that around a stationary point q_* or for the initial iterate q_0 (which in this paper is assumed to be deterministic), the norm of the noise is not heavy-tailed. When $x = q_k$ is a random variable this is no longer the case. This is in line with e.g [26], in which it is reported that the noise is not heavy-tailed initially.

Lemma B.4. *Under Assumption 5 it holds that*

$$\mathbb{E}\left[\|\nabla F(q_k) - \nabla f(q_k,\xi_k)\|^2\right] < +\infty, \tag{45}$$

where $\{q_k\}_{k\geq 0}$ *is determined by* (10).

Proof. By Lemma B.2 and the triangle inequality it holds that

$$\|\nabla F(q_k) - \nabla f(q_k,\xi_k)\| \leq 2NL_1\left(F(q_k) - F_*\right) + \frac{L_0}{L_1} + \|\nabla F(q_k)\|$$
$$\leq 2L_1(N+1)\left(F(q_k) - F_*\right) + \frac{2L_0}{L_1},$$

where we have used Lemma B.1 to bound the $\|\nabla F(q_k)\|$-term in the second inequality. We can use this to bound one of the powers in the expectation of (45):

$$\mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k,\xi_k)\|^2\right] \leq \mathbb{E}_{\xi_k}\left[\|\nabla F(q_k) - \nabla f(q_k,\xi_k)\|\right]\left(2L_1(N+1)\left(F(q_k) - F_*\right) + \frac{2L_0}{L_1}\right)$$
$$\leq \sigma\cdot\left(2L_1(N+1)\left(F(q_k) - F_*\right) + \frac{2L_0}{L_1}\right),$$

where we have used Assumption 5.ii), along with the fact that q_k is $\mathcal{F}_k$-measurable (hence we may take the $F(q_k)$-term outside of the conditional expectation) while ξ_k is independent of $\mathcal{F}_k$. Taking the total expectation of this, we find that

$$\mathbb{E}\left[\|\nabla F(q_k) - \nabla f(q_k,\xi_k)\|^2\right] \leq \sigma\cdot\left(2L_1(N+1)\mathbb{E}\left[F(q_k) - F_*\right] + \frac{2L_0}{L_1}\right).$$

By Theorem 3.7 we have that $\sup_{k\geq 0} \mathbb{E}\left[F(q_k)\right] < \infty$, and thus (45) follows. $\square$

Appendix C. Existence and uniqueness of solutions

When Assumption 3.i) is satisfied, the solutions to (9) are clearly unique and exist for all $t \geq 0$. Under Assumption 4.i), this is not so obvious anymore. Essentially this follows from the coercivity Assumptions 1.i) and 2.iii), and in this section we elaborate on why this is the case. For brevity, let $z = (p, q)$ and let

$$\dot{z} = f(z) \tag{46}$$

denote the system (9). Put $V(z) = H(p,q) - F_* - \varphi_*$, where the infima F_* and φ_* are defined in Assumptions 1 and 2, respectively. Then V is a Lyapunov function for (46) by Lemma 3.12 and, by construction, $V(z) \geq 0$. We also note that by Assumption 1.i) and 2.iii)

$$\lim_{\|z\|\to\infty} V(z) = \infty.$$

By Theorem 10.1 in [63] (see also Theorem 1 in [62]), it follows that the solutions to (46) are *equi-bounded*, in the sense that for a given $B > 0$, there exists a C such that all solutions with initial conditions $\|z_0\| \leq B$ satisfies $\|z(t)\| \leq C$. Thus the solutions to (46) are defined for all $t \geq 0$, compare §23 in [32]. Since $\nabla\varphi$ is Lipschitz continuous and ∇F is locally Lipschitz continuous, the vector field in (9) is locally Lipschitz continuous. Repeated applications of the Picard–Lindelöf theorem, see e.g. [65, Theorem 3.A], therefore shows that the solutions are unique.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

[2] S. Andradóttir. A new algorithm for stochastic optimization. In O. Balci, R.P. Sadowski, and R.E. Nance, editors, *Proceedings of the 1990 Winter Simulation Conference*, pages 364–366, 1990.

[3] M. D. Asic and D. D. Adamovic. Limit points of sequences in metric spaces. *Am. Math. Mon.*, 77(6):613–616, 1970.

[4] A. Barakat, P. Bianchi, W. Hachem, and S. Schechtman. Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance. *Electron. J. Stat.*, 15(2), 2021.

[5] H. Bateman. On dissipative systems and related variational principles. *Phys. Rev.*, 38, 1931.

[6] A. Beck. *First-Order Methods in Optimization.* SIAM, 2017.

[7] M. Benaïm. A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.*, 34(2), 1996.

[8] Michel Benaïm. Dynamics of stochastic approximation algorithms. In J. Azéma, M. Émery, M. Ledoux, and M. Yor, editors, *Séminaire de Probabilités XXXIII.* Springer, 1999.

[9] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.*, 5(2), 1994.

[10] J. Bico, J. Ashmore-Chakrabarty, G. H. McKinley, and H. A. Stone. Rolling stones: The motion of a sphere down an inclined plane coated with a thin liquid film. *Phys. Fluids*, 21(8):082103, 2009.

[11] D. Bideau, F. X. Riguidel, A. Hansen, G. Ristow, X. I. Wu, K. J. Måløy, and M. Ammi. Granular flow: Some experimental results. In K. K. Bardhan, B. K. Chakrabarti, and A. Hansen, editors, *Non-Linearity and Breakdown in Soft Condensed Matter.* Springer, 1994.

[12] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60(2):223–311, 2018.

[13] M. Crawshaw, M. Liu, F. Orabona, W. Zhang, and Z. Zhuang. Robustness to unbounded smoothness of generalized signsgd. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9955–9968. Curran Associates, Inc., 2022.

[14] A. Defazio. Momentum via primal averaging: Theoretical insights and learning rate schedules for non-convex optimization, 2021. arXiv: 2010.00406 [cs.LG].

[15] S. N. Ethier and T. G. Kurtz. *Markov processes – characterization and convergence.* Wiley, 1986.

[16] C. Fang, Z. Lin, and T. Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. *PMLR*, 99:1192–1234, 2019.
[17] M. Faw, L. Rout, C. Caramanis, and S. Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive SGD. In G. Neu and L. Rosasco, editors, *COLT*, volume 195 of *Proceedings of Machine Learning Research*, pages 89–160. PMLR, 2023.
[18] G. Franca, M. I. Jordan, and R. Vidal. On dissipative symplectic integration with applications to gradient-based optimization. *J. Stat. Mech.: Theory Exp.*, 2021.
[19] G. Franca, J. Sulam, D. Robinson, and R. Vidal. Conformal symplectic and relativistic optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.
[20] F. Freise. *On Convergence of the Maximum Likelihood Estimator in Adaptive Designs*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, 2016.
[21] S. Gadat, F. Panloup, and F. Saadane. Stochastic heavy ball. *Electron. J. Stat.*, 12(1), 2018.
[22] P. Glendinning. *Stability, instability and chaos; an introduction to the theory of nonlinear differential equations*. Cambridge University Press, 1994.
[23] H. Goldstein, C.P.Jr. Poole, and J.L. Safko. *Classical Mechanics*. Addison-Wesley, 3rd edition, 2014.
[24] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.
[25] E. Gorbunov, M. Danilova, and A. Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15042–15053. Curran Associates, Inc., 2020.
[26] M. Gurbuzbalaban, U. Simsekli, and L. Zhu. The heavy-tail phenomenon in SGD. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.
[27] J. Kapoor and Harshvardhan. A stochastic extension of Hamiltonian descent methods. URL `https://harshv834.github.io/files/shd_report.pdf`, accessed 9 January 2024, 2021.
[28] A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023.
[29] A. Koloskova, H. Hendrikx, and S. U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning*, 2023.
[30] H. J. Kushner and D. S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Springer, 1st edition, 1978.
[31] H.J. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition, 2003.
[32] J. LaSalle and S. Lefschetz. *Stability by Liapunov's Direct Method with applications*. Academic press, 1961.
[33] B. Leimkuhler and C. Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Springer, 2015.
[34] H. Li, A. Rakhlin, and A. Jadbabaie. Convergence of Adam under relaxed assumptions. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 52166–52196. Curran Associates, Inc., 2023.
[35] S. Livingstone, M. F. Faulkner, and G.O. Roberts. Kinetic energy choice in Hamiltonian/hybrid Monte Carlo. *Biometrika*, 2017.
[36] L. Ljung. *Analysis of Recursive Stochastic Algorithms*. Technical Reports TFRT-7097. Department of Automatic Control, Lund Institute of Technology (LTH), 1976.
[37] C. J. Maddison, D. Paulin, Y. W. Teh, B. O'Donoghue, and A. Doucet. Hamiltonian descent methods, 2018. arXiv:1809.05042 [math.OC].
[38] V.V. Mai and M. Johansson. Stability and convergence of stochastic gradient clipping: Beyond Lipschitz continuity and smoothness. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 7325–7335. PMLR, 2021.
[39] T. Mikolov. *Statistical language models based on neural networks*. PhD thesis, Brno University of Technology, 2013.
[40] Y. Nesterov. *Lectures on Convex Optimization*. Springer, 2018.
[41] A.J. Owens and D.L. Filkin. Efficient training of the backpropagation network by solving a system of stiff ordinary differential equations. In *International 1989 Joint Conference on Neural Networks*, volume 2, pages 381–386, 1989.
[42] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training neural networks. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28(3), 2013.
[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
[44] A.L. Peressini, F.E. Sullivan, and J.J. Uhl, Jr. *The Mathematics of Nonlinear Programming*. Springer, 1993.
[45] B. Poljak. A general method for solving extremum problems. *Sov. Math., Dokl.*, 8(3), 1967.

[46] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Comp. Math. Math. Phys.*, 4, 1964.

[47] H. Robbins and S. Monro. A stochastic approximation algorithm. *Ann. Math. Stat.*, 22(3), 1951.

[48] H. Robbins and D. Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In J.S. Rustagi, editor, *Optimizing Methods in Statistics*. Academic Press, 1971.

[49] T. Rolski, H. Schmidli, V. Schmidt, and J.L. Teugels. *Stochastic Processes for Insurance and Finance*. Wiley Series in Probability and Statistics. Wiley, 2009.

[50] A. Sard. The measure of the critical values of differentiable maps. *Bull. Amer. Math. Soc.*, 48(12):883–890, 1942.

[51] O. Sebbouh, R. M. Gower, and A. Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In M. Belkin and S. Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134. PMLR, 2021.

[52] J. Sherman. Nearly maximum flows in nearly linear time. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 263–269, 2013.

[53] A.N. Shiryaev. *Probability-1*. Graduate Texts in Mathematics. Springer, 3rd edition, 2016.

[54] G. Stoltz and Z. Trstanova. Langevin dynamics with general kinetic energies. *Multiscale Model. Sim.*, 16(2), 2018.

[55] M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. *Stat*, 9(1), 2020.

[56] B. Wang, H. Zhang, Z. Ma, and W. Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In G. Neu and L. Rosasco, editors, *COLT*, volume 195 of *Proceedings of Machine Learning Research*. PMLR, 2023.

[57] B. Wang, H. Zhang, Z. Ma, and W. Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In G. Neu and L. Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*. PMLR, 2023.

[58] H. Whitney. A function not constant on a connected set of critical points. *Duke Math. J.*, 1(4):514–517, 1935.

[59] S. Wiggins. *Introduction to applied non-linear dynamics and chaos*. Springer New York, 2003.

[60] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[61] D.E. Wolf, F. Radjai, and S. Dippel. Dissipation in granular materials. *Philos. Mag. B*, 77(5):1413–1425, 1998.

[62] T. Yoshizawa. Liapunov's function and boundedness of solutions. *Funkcialaj Ekvacioj*, 2, 1959.

[63] T. Yoshizawa. *Stability Theory by Liapunov's second method*. The mathematical society of Japan, 1966.

[64] E. Zeidler. *Nonlinear Functional Analysis and Its Applications: III: Variational Methods and Optimization*. Springer, 1985.

[65] E. Zeidler. *Nonlinear functional analysis and its applications. I*. Springer, 1986. Fixed-point theorems.

[66] B. Zhang, J. Jin, C. Fang, and L. Wang. Improved analysis of clipping algorithms for non-convex optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15511–15521. Curran Associates, Inc., 2020.

[67] J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.