WHAT IS THE LONG-RUN DISTRIBUTION OF STOCHASTIC GRADIENT DESCENT? A LARGE DEVIATIONS ANALYSIS

WAÏSS AZIZIAN[c,*], FRANCK IUTZELER[♯], JÉRÔME MALICK[*], AND PANAYOTIS MERTIKOPOULOS[◇]

Abstract. In this paper, we examine the long-run distribution of stochastic gradient descent (SGD) in general, non-convex problems. Specifically, we seek to understand which regions of the problem's state space are more likely to be visited by SGD, and by how much. Using an approach based on the theory of large deviations and randomly perturbed dynamical systems, we show that the long-run distribution of SGD resembles the Boltzmann–Gibbs distribution of equilibrium thermodynamics with temperature equal to the method's step-size and energy levels determined by the problem's objective and the statistics of the noise. In particular, we show that, in the long run, (*a*) the problem's critical region is visited exponentially more often than any non-critical region; (*b*) the iterates of SGD are exponentially concentrated around the problem's minimum energy state (which does not always coincide with the global minimum of the objective); (*c*) all other connected components of critical points are visited with frequency that is exponentially proportional to their energy level; and, finally (*d*) any component of local maximizers or saddle points is "dominated" by a component of local minimizers which is visited exponentially more often.

1. Introduction

Even though stochastic gradient descent (SGD) has been around for more than 70 years [61], it is still the method of choice for training a wide array of modern machine learning architectures – from large language models to reinforcement learning and recommender systems. This phenomenal success is largely owed to the method's simplicity: given a smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$ and the associated optimization problem

$$\min_{x\in\mathbb{R}^d} \; f(x) \tag{Opt}$$

the SGD algorithm is given by the simple update rule

$$x_{n+1} = x_n - \eta \hat{g}_n \tag{SGD}$$

where $\eta > 0$ is the method's step-size and $\hat{g}_n$, $n = 0, 1, \dots$ is a stochastic gradient of f at x_n.

By virtue of its wide applicability, (SGD) and its variants have been studied extensively in the literature, for both convex and non-convex objectives. In the non-convex case (which is the most relevant setting for machine learning), the basic, no-frills guarantees of (SGD) boil down to bounds of the form $\mathbb{E}\big[\sum_{k=0}^{n}\|\nabla f(x_k)\|^2\big] = \mathcal{O}(\sqrt{n})$ provided that η has been

[c] Corresponding author.

[*] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France.

[♯] Institut de Mathématiques de Toulouse, Université de Toulouse, CNRS, UPS, 31062, Toulouse, France.

[◇] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France.

E-mail addresses: waiss.azizian@univ-grenoble-alpes.fr, franck.iutzeler@math.univ-toulouse.fr, jerome.malick@cnrs.fr, panayotis.mertikopoulos@imag.fr.

2020 *Mathematics Subject Classification.* Primary 90C15, 90C26, 60F10; secondary 90C30, 68Q32.

Key words and phrases. Stochastic gradient descent; Freidlin–Wentzell theory; large deviations; Boltzmann–Gibbs distribution.

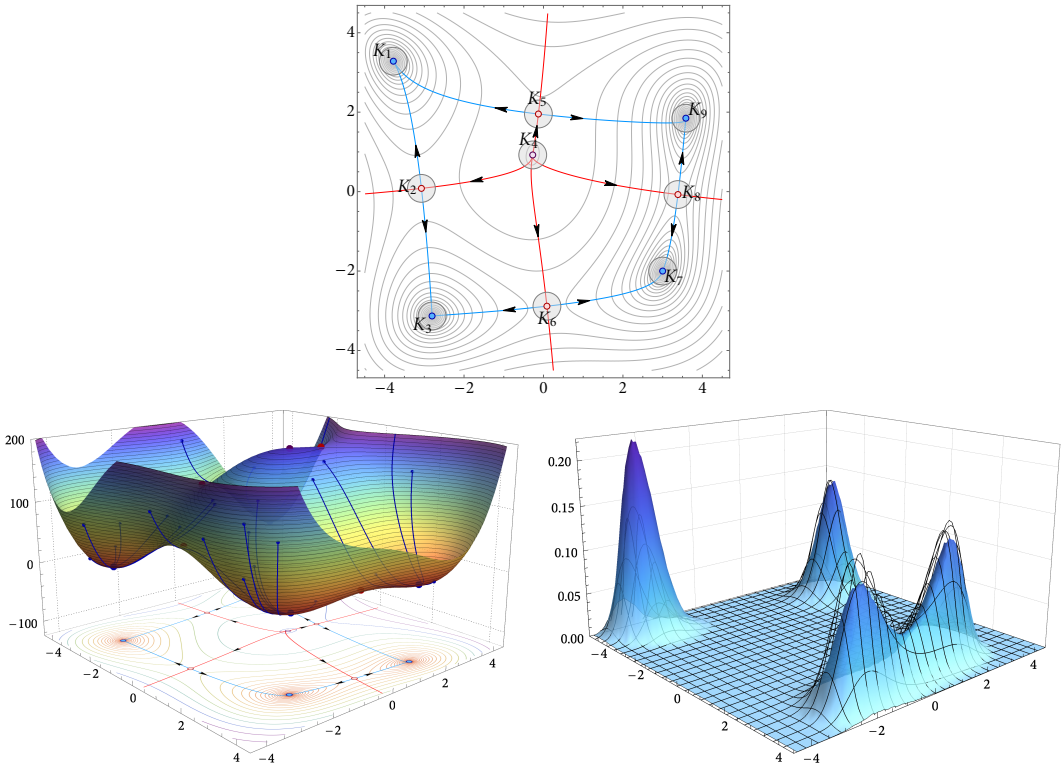

Figure 1: Graphical illustration of Theorems 1–4 for the Himmelblau test function $f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$. The figure to the left depicts the loss landscape of f with several (deterministic) orbits of the gradient flow of f superimposed for visual convenience. The figure in the middle highlights the 9 critical components of f as well as the "most likely" transitions between them: $\mathcal{K}_1$, $\mathcal{K}_3$, $\mathcal{K}_7$ and $\mathcal{K}_9$ are minimizers (light blue), $\mathcal{K}_5$ is a global maximum (light purple), and the rest are saddle points (light red). The figure to the right illustrates the long-run distribution of 1000 samples of (SGD) run with $\eta = 0.01$ over a horizon of 2×10^4 iterations: the density landscape represents the observed distribution, while the superimposed wireframe indicates our theoretical prediction.

chosen accordingly [37]. This guarantee suggests that the sequence x_n eventually spends all but a vanishing fraction of time near regions where ∇f is small, but it does not answer *where* (SGD) ultimately settles down. In particular, the following crucial question remains open:

Which critical points of f (or components thereof) are more likely to be observed in the long run – and by how much?

This question is notoriously difficult because the loss landscape of f can be exceedingly complicated – especially in deep learning problems with hundreds of millions (or even billions) of parameters. Starting with the negative, f may contain a number of spurious saddle points that is exponentially larger than the number of local minima, and the function values associated with worst-case saddle points may be considerably worse than those associated with worst-case local minima [10]. On the flip side, a more positive answer is provided by the literature on the avoidance of saddle-points where, under different assumptions for the method's step-size and the structure of f, it has been shown that the time spent by x_n in the

vicinity of strict saddle points is (vanishingly) small; for a representative – but, by necessity, incomplete – list of results, cf. [2, 4, 21, 26, 27, 31, 39, 52, 58, 69] and references therein.

Now, even though the above justifies the informal mantra that "SGD avoids saddle points", it does not answer *which* critical regions of f are most likely to be observed in the long run, and by how much. This question has attracted significant interest in deep learning, but the matter remains poorly understood: on the one hand, some works have shown that, in certain stylized deep net models, most local minimizers are concentrated in an exponentially narrow band of the problem's global minimum [8, 33]; on the other hand, empirical studies suggest that, even in this case, the long-run distribution of (SGD) may not be adequately captured by the shape of the problem's loss function [47].

Our contributions. Our goal in this paper is to quantify the long-run distribution of (SGD) in the most general manner possible. To do so, we take an approach based on the theory of large deviations [12] and randomly perturbed dynamical systems [20, 35], which enables us to estimate the probability of "rare events" (such as x_n moving against the gradient flow of f for a protracted period of time). This allows us to characterize the events that occur with high probability and establish the following hierarchy of results (stated formally as Theorems 1–4 in Section 3):

(1) In the long run, the critical region of f is visited exponentially more often than any non-critical region.

(2) The iterates of (SGD) are concentrated with exponentially high probability in the vicinity of a region that minimizes a certain "energy functional" which depends on f and the statistics of the noise in (SGD). Importantly, the ground state of this functional *does not* necessarily coincide with the global minimum of f.

(3) Among the remaining connected components of critical points, each component is visited with frequency which is exponentially proportional to its energy, according to the Boltzmann–Gibbs distribution of statistical physics with temperature equal to the method's step-size.

(4) Every connected component of non-minimizing critical points of f – i.e., local maximizers or saddle points – is "dominated" by a component of local minimizers that is visited exponentially more often.

Finally, we derive an explicit characterization of the invariant measure of (SGD) under Gaussian noise and other noise models motivated by deep learning considerations.

Taken together, these properties resemble those of a canonical ensemble in statistical physics: in a sense, each connected component of critical points can be seen as a "state" of a statistical ensemble, and the step-size of (SGD) plays the role of the system's (fixed) temperature, which determines how easy it is to transit from one component to another. We find this analogy particularly appealing as it provides a way of connecting ideas from equilibrium thermodynamics to the long-run behavior of (SGD).

Related work. The main approaches used in the literature to examine the long-run distribution of (SGD) hinge on the study of a limiting stochastic differential equation (SDE), typically associated with (a version of) the discrete Langevin dynamics or a diffusion approximation of (SGD).

Starting with the former, Raginsky et al. [59] examined the law of the stochastic gradient Langevin dynamics (SGLD), a variant of (SGD) with injected Gaussian noise of variance $2\eta/\beta$ for some inverse temperature parameter $\beta > 0$ (see also [18, 40] for some recent follow-ups in this direction). Raginsky et al. [59] first showed that SGLD closely tracks an associated diffusion process over finite time intervals; the law of this diffusion was then shown

to converge to the Gibbs measure $\exp(-\beta f)/\int \exp(-\beta f)$ at a geometric rate, fast enough to ensure the convergence of the discrete-time dynamics to the same measure.

(SGD) can be recovered in the context of SGLD by setting the inverse temperature parameter to $\beta \propto 1/\eta$. Unfortunately however, the convergence rate of the SDE to its invariant distribution is exponential in β, so it is too slow to compensate for the discretization error in this case. As a result, the bounds between the discrete dynamics and the invariant measure of the limiting diffusion become vacuous in the case of (SGD) – and similar considerations apply to the related work of Majka et al. [51].[1]

Another potential approach to studying the long-run distribution of (SGD) consists of approximating its trajectories via the solutions of a limiting SDE. A key contribution here was provided by the work of Li et al. [42, 43] who showed that the tracking error between the iterates of (SGD) and the solution of a certain SDE becomes vanishingly small in the limit $\eta \to 0$ over any *finite* time interval. However, in contrast to the Langevin case, the convergence speed of the induced stochastic modified equation (SME) to its invariant measure degrades exponentially as $\eta \to 0$ [17], rendering this approach moot for a global description of the invariant measure of (SGD). Though this strategy has been refined, either in the vicinity of global minimizers [6, 45] or in regions where f is locally strongly convex [19, 41], this diffusion approach still fails to capture the long run behavior of (SGD) in general non-convex settings.

Nevertheless, the limiting SDE still provides valuable insights into certain aspects of the dynamics of (SGD). In particular, a fruitful thread in the literature [28, 30, 56, 68] has sought to estimate the escape rates of the approximating diffusion from local minimizers through this approach. Interestingly, these works use some elements of the continuous-time Freidlin–Wentzell theory [20], which is also the point of departure of our paper. That being said, even though these results demonstrate how the structure of the objective function and of the noise locally affect the dynamics of the SDE in a basin, they provide no information on the long-run behavior of the (discrete-time) dynamics of (SGD); for a more technical discussion, see Appendix A.

One last approach which has gained increased attention in the literature is that of Dieuleveut et al. [13] and Lu et al. [49] who study (SGD) as a discrete-time Markov chain. This allowed [13, 49] to derive conditions under which (SGD) is (geometrically) ergodic and, in this way, to quantify the bias of the invariant measure under global growth conditions, i.e., the distance to the global minimum of f. Building further on this perspective, Gürbüzbalaban et al. [23], Hodgkinson & Mahoney [25] and Pavasovic et al. [57] showed that, under general conditions, the asymptotic distribution of the iterates of (SGD) is heavy-tailed; however, these results only describe the distribution of (SGD) near infinity, and they provide no information on which critical regions of f are more likely to be observed. Again, we provide some more details on this in Appendix A.

Our approach and techniques. The linchpin of our approach is the theory of large deviations of Freidlin & Wentzell [20] for Markov processes, originally developed for diffusion processes in continuous time, and subsequently extended to subsampling in discrete time by Kifer [35]; see also [11, 22] and [15, 16] for applications to stochastic approximation. However, the starting point of all these works is the study of continuous-time diffusions on closed manifolds; as far as we are aware, our paper provides the first extension of the theory of

[1] In more detail, the error term in Raginsky et al. [59, Eq. (3.1)] can no longer be controlled if η is small. Similarly, the constants in the geometric convergence rate guarantees of Majka et al. [51, Theorems 2.1 and 2.5] would degrade as $\exp(-\Omega(1/\eta))$; as a result, the associated discretization errors would be of the order of $\exp(\Omega(1/\eta)))$, which cannot be controlled for small η.

Freidlin & Wentzell to discrete-time systems that evolve over unbounded domains, and with a general – possibly discrete – noise profile.

One of the key challenges that we need to overcome is that most of the potentials introduced in [20, 35] become drastically less regular in our context; we remedy this issue by refining the analysis and carefully studying the structure of the attractors of (SGD). This allows us to salvage enough regularity and show that (SGD) spends most of its time near its attractors (this is achieved by developing suitable tail-bounds for the time spent away from critical points) and, ultimately, to estimate the transition probabilities of (SGD) between different connected components of critical points. This involves a series of novel mathematical tools and techniques, which we detail in Appendix D.

2. Preliminaries and blanket assumptions

2.1. **Standing assumptions.** In this section, we describe our assumptions for the objective function f of (Opt) and the black-box oracle providing gradient information for (SGD). We begin with the former, writing throughout $\mathcal{X} \coloneqq \mathbb{R}^d$ for the domain of f.

Assumption 1. The objective function $f\colon \mathcal{X} \to \mathbb{R}$ satisfies the following conditions:

(a) *Coercivity:* $f(x) \to \infty$ as $\|x\| \to \infty$.

(b) *Smoothness:* f is C^2-differentiable and its gradient is β-Lipschitz continuous, that is,

$$\|\nabla f(x') - \nabla f(x)\| \leq \beta \|x' - x\| \quad \text{for all } x, x' \in \mathcal{X}\,. \tag{LG}$$

(c) *Critical set regularity:* The critical set

$$\operatorname{crit} f \coloneqq \{x \in \mathcal{X} : \nabla f(x) = 0\} \tag{1}$$

of f consists of a finite number of smoothly connected components $\mathcal{K}_i$, $i = 1, \dots, K$.

These requirements are fairly standard in the literature: Assumption 1(a) guarantees that (Opt) admits a solution and rules out infima at infinity (such as $f(x) = e^{-x^2}$); Assumption 1(b) is a bare-bones regularity requirement for the analysis of gradient methods; and, finally, Assumption 1(c) serves to exclude objectives with anomalous critical sets (e.g., exhibiting kinks or other non-smooth features), so it is also quite mild from an operational standpoint.[2]

Regarding the gradient input to (SGD), we will assume throughout that the optimizer has access to a *stochastic first-order oracle* (SFO), that is, a black-box mechanism returning a stochastic estimate of the gradient of f at the point of interest. Formally, when queried at $x \in \mathcal{X}$, an SFO returns a random vector of the form

$$\mathsf{G}(x; \omega) = \nabla f(x) + \mathsf{Z}(x; \omega) \tag{SFO}$$

where

(a) ω is a random seed drawn from a compact subset Ω of $\mathbb{R}^m$ based on some (complete) probability measure $\mathbb{P}$.[3]

(b) $\mathsf{Z}(x; \omega)$ is an umbrella error term capturing all sources of noise and randomness in the oracle.

This oracle model is sufficiently flexible to account for all established versions of (SGD) in the literature, including minibatch SGD (where ω represents the sampled minibatch), noisy gradient descent (where the optimizer may artificially inject noise in the process to enhance convergence), and Langevin Monte Carlo methods.

[2] This last requirement can be replaced by positing for example that f is definable in terms of some semi-algebraic / o-minimal structure, see e.g., [9, 63] and Remark B.1 in Appendix B.

[3] The specific form of Ω is not important; in practice, random seeds from a target distribution are often generated by inverse transform sampling from $[0, 1]^m$.

With all this in mind, we make the following blanket assumptions for (SFO):

Assumption 2. The error term $\mathsf{Z}\colon \mathcal{X} \times \Omega \to \mathbb{R}^d$ of (SFO) satisfies the following properties:

(*a*) *Properness:* $\mathbb{E}[\mathsf{Z}(x;\omega)] = 0$ and $\operatorname{cov}(\mathsf{Z}(x;\omega)) \succ 0$ for all $x \in \mathcal{X}$.

(*b*) *Smooth growth:* $\mathsf{Z}(x;\omega)$ is C^2-differentiable and satisfies the growth condition

$$\sup_{x,\omega} \frac{\|\mathsf{Z}(x;\omega)\|}{1+\|x\|} < \infty\,. \tag{2}$$

(*c*) *Sub-Gaussian tails:* The tails of Z are bounded as

$$\log \mathbb{E}\Big[e^{\langle p, \mathsf{Z}(x;\omega)\rangle}\Big] \leq \frac{1}{2}\sigma_\infty^2 \|p\|^2 \tag{3}$$

for some $\sigma_\infty > 0$ and all $p \in \mathbb{R}^d$.

Assumption 2(*a*) is standard in the literature and ensures that the gradient noise in (SGD) has zero mean and does not vanish identically at any $x \in \mathcal{X}$; this requirement in particular plays a crucial role in several incarnations of noisy gradient descent that have been proposed to effectively escape saddle points of f [4, 21, 31, 58]. Assumption 2(*b*) is a bit more technical but otherwise simply serves to impose a limit on how large the noise may grow as $\|x\| \to \infty$. Finally, Assumption 2(*c*) is also widely used in the literature: while not as general as the (possibly fat-tailed) finite variance assumption $\mathbb{E}[\|\mathsf{Z}(x;\omega)\|^2] \leq \sigma_\infty^2$, it allows much finer control of the stochastic processes involved, leading in turn to more explicit and readily interpretable results. We only note here that Assumption 2(*c*) can be relaxed further by allowing the variance proxy σ_∞^2 of $\mathsf{Z}(x;\omega)$ to depend on x, possibly diverging to infinity as $\|x\| \to \infty$. To streamline our presentation, we defer the general case to the appendix.

Our last blanket requirement is a stability condition ensuring that the signal-to-noise ratio of (SFO) does not become too small at infinity. We formalize this as follows:

Assumption 3. The signal-to-noise ratio of G is bounded as

$$\liminf_{\|x\|\to\infty} \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2} > 16 \log 6 \cdot d\,. \tag{4}$$

Assumption 3 is a technical requirement needed to establish a series of concentration bounds later on, and the specific value of the lower bound serves to facilitate some computations later on. In practice, ∇f is often norm-coercive – i.e., $\|\nabla f(x)\| \to \infty$ as $\|x\| \to \infty$ – so this assumption is quite mild. Note also that Assumption 3, as Assumption 2(*c*), can be extended to the case where the variance proxy σ_∞^2 of G depends on x, possibly blowing up at infinity; we postpone the relevant details to Appendix B.2.

Putting together all of the above, the SGD algorithm can be written in abstract recursive form as

$$x^+ \leftarrow x - \eta\, \mathsf{G}(x;\omega)\,. \tag{5}$$

Thus, given a (possibly random) initialization $x_0 \in \mathcal{X}$ and an i.i.d. sequence of random seeds $\omega_n \in \Omega$, $n = 0, 1, \dots$, the iterates x_n of (SGD) are obtained by taking $\hat{g}_n \leftarrow \mathsf{G}(x_n;\omega_n)$ and iterating $n \leftarrow n+1$ ad infinitum. To streamline notation, we will write $\mathbb{P}_{x_0}$ for the law of x_n starting at x_0, and we will refer to it as the law of (SGD).

2.2. **Discussion of the assumptions.** To illustrate the generality of our assumptions, we briefly consider here the example of the regularized empirical risk minimization problem

$$f(x) = \frac{1}{n}\sum_{i=1}^n \ell(x;\xi_i) + \frac{\lambda}{2}\|x\|^2 \tag{6}$$

where ξ_i, $i = 1, \dots, n$, are the training data of the model, $\ell(x;\xi)$ represents the loss of the model x on the data point ξ, and $\lambda > 0$ is a regularization parameter. In this case, we have

$$\mathsf{Z}(x;\omega) = \nabla\ell(x;\xi_\omega) - \frac{1}{n}\sum_{i=1}^{n} \nabla\ell(x;\xi_i) \tag{7}$$

where ω is sampled uniformly at random from $\{1, \dots, n\}$.

If ℓ is C^2-differentiable, Lipschitz continuous and smooth – see e.g., [52] and references therein – Assumption 1 and (a) and (b) are satisfied automatically. The error term $\mathsf{Z}(x;\omega)$ is also uniformly bounded so Assumption 2 and (b) and (c) are likewise verified (see e.g., Wainwright [65, Ex. 2.4]). Finally, we have $\|\nabla f(x)\| = \mathcal{O}(\lambda\|x\|)$, so Assumption 3 also holds. Thus, this setting covers two wide classes of examples: linear models with non-convex losses [18, 49] and smooth neural networks with normalization layers [44].[4]

3. Analysis and results

3.1. **Invariant measures.** The overarching objective of our paper is to understand the statistics of the limiting behavior of (SGD). To that end, our point of departure will be the *mean occupation measure* of x_n, defined here as

$$\mu_n(\mathcal{B}) = \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}\{x_k \in \mathcal{B}\}\right] \tag{8}$$

for every Borel $\mathcal{B} \subseteq \mathcal{X}$. In words, $\mu_n(\mathcal{B})$ simply measures the mean fraction of time that x_n has spent in $\mathcal{B}$ up to time n, so the long-run statistics of (SGD) can be quantified by the limiting distribution $\lim_{n\to\infty} \mu_n$.

If x_n is ergodic, μ_n converges weakly to some measure μ_∞, known as the *invariant measure* of x_n.[5] Referring to the abstract representation (5) of (SGD), "invariance" simply means here that μ_∞ satisfies the defining property

$$x \sim \mu_\infty \implies x^+ \sim \mu_\infty \tag{9}$$

i.e., μ_∞ is stationary under (SGD). Note however that ergodicity generally requires that the noise has a density part [49]. More generally, even if x_n is not ergodic, any (weak) limit point of μ_n must still satisfy the invariance property (9) [14, 24], so this will be our principal figure of merit.[6]

More precisely, our goal will be to quantify the long-run concentration of probability mass near the components $\mathcal{K}_i$ of $\operatorname{crit} f$: in particular, since each $\mathcal{K}_i$ generically has Lebesgue measure zero, we will seek to estimate the probability mass $\mu_\infty(\mathcal{U}_i)$ where μ_∞ is invariant under (SGD) and $\mathcal{U}_i$ is a sufficiently small neighborhood of $\mathcal{K}_i$ – typically a δ-neighborhood of the form $\mathcal{U}_i \equiv \mathcal{U}_i(\delta) := \{x \in \mathcal{X} : \operatorname{dist}(\mathcal{K}_i, x) < \delta\}$ in the limit $\delta \to 0$. Doing this will allow us to determine the probability that (SGD) is concentrated in the long run near one critical component or another, as well as the degree of this concentration.

---

[4] Our assumptions can also be linked to the notion of *dissipativity*, which is standard in Markov chain and sampling literature, see e.g., [18, 40, 49, 50, 59]. Considering the gradient oracle obtained by sampling minibatches of size B, this setting fits into our framework with the relaxed version of Assumption 2 in Appendix B provided B is chosen large enough, see Appendix B.2.

[5] Weak convergence means here that $\lim_{n\to\infty} \int \varphi \, d\mu_n = \int \varphi \, d\mu_\infty$ for every bounded continuous function $\varphi\colon \mathcal{X} \to \mathbb{R}$.

[6] Existence always holds in our setting, cf. Lemma D.16 in Appendix D.3.

3.2. **A large deviation principle for SGD.** Our strategy to achieve this will be to estimate the long-run rates of transition between different regions of $\mathcal{X}$ under (SGD). Our first step in this regard will be to establish a *large deviation principle* (LDP) for the process x_n, $n = 0, 1, \dots$, in the spirit of the general theory of Freidlin & Wentzell [20]. This will in turn allow us to quantify the probability of "rare events" in (SGD) – e.g., moving against the gradient flow of f for a protracted period of time – and it will play a crucial role in the sequel.

Now, since the statistics of (SGD) are determined by those of (SFO), we begin by considering the *cumulant-generating functions* (CGFs) of Z and G, viz.

$$K_{\mathsf{Z}}(x,p) \coloneqq \log \mathbb{E}[\exp(\langle p, \mathsf{Z}(x;\omega)\rangle)] \tag{10a}$$

$$\begin{aligned} K_{\mathsf{G}}(x,p) &\coloneqq \log \mathbb{E}[\exp(\langle p, \mathsf{G}(x;\omega)\rangle)] \\ &= K_{\mathsf{Z}}(x,p) + \langle \nabla f(x), p\rangle \end{aligned} \tag{10b}$$

where $x \in \mathcal{X}$, $p \in \mathbb{R}^d$, and $\langle p, v\rangle$ denotes the standard bilinear pairing between $p \in \mathbb{R}^d$ and $v \in \mathcal{X}$. To state our results, we will also require the associated *Lagrangians*

$$\mathcal{L}_{\mathsf{Z}}(x,v) \coloneqq K_{\mathsf{Z}}^*(x,-v) \tag{11a}$$

$$\mathcal{L}_{\mathsf{G}}(x,v) \coloneqq K_{\mathsf{G}}^*(x,-v) = \mathcal{L}_{\mathsf{Z}}(x, v + \nabla f(x)) \tag{11b}$$

where "$*$" denotes convex conjugation with respect to p in $K_{\mathsf{Z}}(x,p)$ and $K_{\mathsf{G}}(x,p)$.

The importance of the Lagrangian functions (11) is that they provide a large deviation principle for Z and G. Namely, to leading order in n (and ignoring boundary effects), we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=0}^{n} \mathsf{G}(x;\omega_n) \in \mathcal{B}\right) \sim \exp\left(-n \inf_{v\in\mathcal{B}} \mathcal{L}_{\mathsf{G}}(x,v)\right) \tag{12}$$

for every Borel $\mathcal{B} \in \mathbb{R}^d$ (and likewise for Z and $\mathcal{L}_{\mathsf{Z}}$), so the long-run statistics of the process $S_n = \sum_{k=0}^{n-1} \mathsf{G}(x;\omega_k)$ are fully determined by $\mathcal{L}_{\mathsf{G}}$ [12]. In this regard, $\mathcal{L}_{\mathsf{G}}$ plays the role of a "rate function" for S_n and quantifies the rate of occurrence of "rare events" in this context [12].

Going back to (SGD), we have $x_n = x_0 - \eta \sum_{k=0}^{n-1} \mathsf{G}(x_k;\omega_k)$, so a promising way to understand the occupation measure of x_n would be to try to derive a large deviation principle for x_n starting from (12). Unfortunately however, in contrast to $S_n = \sum_{k=0}^{n-1} \mathsf{G}(x;\omega_k)$, this is not possible because (12) concerns i.i.d. samples drawn at a fixed point $x \in \mathcal{X}$, while the iterates of x_n are highly auto-correlated. Instead, inspired by the theory of Freidlin & Wentzell [20] for randomly perturbed dynamical systems, we will encode the *entire* trajectory x_n of (SGD) as a point in some infinite-dimensional space of curves, and we will derive a large deviation principle for (SGD) directly in this space.

To make this idea precise, we first require a continuous-time surrogate for the sequence of iterates of (SGD). Concretely, writing $\tau_n = n\eta$ for the "effective time" that has elapsed up to the n-th iteration of (SGD), we define the continuous-time interpolation of x_n as the piecewise affine curve

$$X(t) = x_n + \frac{t - \tau_n}{\eta}(x_{n+1} - x_n) \tag{13}$$

for all $n = 0, 1, \dots$, and all $t \in [\tau_n, \tau_{n+1}]$. The resulting curve is continuous by construction, so, for embedding purposes, we will consider the ambient spaces of continuous curves truncated at some finite $T \geq 0$:

$$\mathcal{C}_T \coloneqq \mathcal{C}([0,T], \mathcal{X}) \tag{14a}$$

$$\mathcal{C}_T(x) \coloneqq \{\gamma \in \mathcal{C}_T : \gamma(0) = x\} \tag{14b}$$

$$\mathcal{C}_T(x,x') \coloneqq \{\gamma \in \mathcal{C}_T : \gamma(0) = x, \gamma(T) = x'\}. \tag{14c}$$

With these preliminaries in hand, and in analogy with the Lagrangian formulation of classical mechanics, we define the (normalized) "*action functional*" of $\mathcal{L}_{\mathsf{G}}$ as

$$\mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}_{\mathsf{G}}(\gamma(t), \dot{\gamma}(t))\, dt \tag{15}$$

for all $\gamma \in \mathcal{C}_T$ and with the convention $\mathcal{S}_T[\gamma] = \infty$ if γ is not absolutely continuous. In a certain sense (to be made precise below), the functional $\mathcal{S}_T[\gamma]$ is a "measure of likelihood" for the curve γ, with lower values indicating higher probabilities. Accordingly, by leveraging the so-called "least action principle" [12, 20, 36], it is possible to establish the following large deviation principle for (SGD):

Proposition 1. *Fix a time horizon $T > 0$, tolerance margins $\varepsilon, \delta > 0$, and an action level $s > 0$. In addition, write*

$$\Gamma_T(x_0; s) \coloneqq \{\gamma \in \mathcal{C}_T(x_0) : \mathcal{S}_T[\gamma] \leq s\} \tag{16}$$

for the space of continuous curves starting at x_0 and with action at most s. Then, for all sufficiently small η, we have

$$\begin{aligned} &\mathbb{P}_{x_0}\left(\sup_{0\leq t\leq T} \|X(t) - \gamma(t)\| < \delta\right) \\ &\qquad \geq \exp\left(-\frac{\mathcal{S}_T[\gamma] + \varepsilon}{\eta}\right) \quad \textit{for all } \gamma \in \Gamma_T(x_0; s) \end{aligned} \tag{17a}$$

and, in addition,

$$\begin{aligned} &\mathbb{P}_{x_0}\left(\sup_{0\leq t\leq T} \|X(t) - \gamma(t)\| > \delta \textit{ for all } \gamma \in \Gamma_T(x_0; s)\right) \\ &\qquad \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right). \end{aligned} \tag{17b}$$

In words, Proposition 1 states that (a) the linear interpolation $X(t)$ of x_n stays close to low-action trajectories with probability that is exponentially large in their action value; and (b) the probability that $X(t)$ strays far from said trajectories is exponentially small in their action value. This is, in fact, the first rung in a hierarchy of large deviation principles that ultimately quantify the probability of rare events in (SGD); because these results are fairly technical to set up and prove (and not required for stating our main result), we defer the relevant discussion and proofs to Appendix C.

3.3. **Transition costs and the quasi-potential.** Now, in view of the characterization (17a) and (17b) of "rare trajectories" of (SGD), we will seek to derive below an analogous characterization for the "typical trajectories" of (SGD) in terms of $\mathcal{S}$. To do this, we will associate a certain *transition cost* to each pair of components $\mathcal{K}_i$, $\mathcal{K}_j$ of $\operatorname{crit} f$, and we will use these costs to quantify how likely it is to observe x_n near a component of critical points of f.

These costs are defined as follows: First, following Freidlin & Wentzell [20], define the *quasi-potential* between two points $x, x' \in \mathcal{X}$ as

$$B(x, x') \coloneqq \inf\{\mathcal{S}_T[\gamma] : \gamma \in \mathcal{C}_T(x, x'), T \in \mathbb{N}\} \tag{18}$$

and the corresponding quasi-potential between two sets $\mathcal{K}, \mathcal{K}' \subseteq \mathcal{X}$ as

$$B(\mathcal{K}, \mathcal{K}') \coloneqq \inf\{B(x, x') : x \in \mathcal{K}, x' \in \mathcal{K}'\}. \tag{19}$$

By construction, $B(x, x')$ is the action value of the "most probable" path from x to x', so it can be interpreted as the "action cost" of moving from x to x'. Accordingly, to capture

the difficulty of x_n leaving the vicinity of a given component of crit f and wandering off to another, we will consider the *cost matrix*

$$B_{ij} \coloneqq B(\mathcal{K}_i, \mathcal{K}_j) \tag{20}$$

where $\mathcal{K}_i, \mathcal{K}_j$, $i, j = 1, \dots, K$, are any two components of critical points of f.

As we mentioned before, the transitions between components of crit f play a crucial role in our analysis because this is where x_n spends most of its time. To characterize the structure of these transitions more precisely, it will be convenient to encode them in a complete weighted directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which we call the *transition graph* of (SGD), and which is defined as follows:

(a) The vertex set of $\mathcal{G}$ is $\mathcal{V} = \{1, \dots, K\}$, i.e., $\mathcal{G}$ has one vertex per component of critical points of f.

(b) The edge set of $\mathcal{G}$ is $\mathcal{E} = \{(i, j) : i, j = 1, \dots, K, i \neq j\}$, i.e., $\mathcal{G}$ has an edge per pair of components of crit f.

(c) The weight of the directed edge $(i, j) \in \mathcal{E}$ is B_{ij}.

To avoid degenerate cases, we will make the following assumption for the problem's cost matrix:

Assumption 4. $B_{ij} < \infty$ for all $i, j = 1, \dots, K$.

This assumption is purely technical and mainly serves to streamline our presentation and avoid complicated statements involving non-communicating classes of the cost matrix B_{ij}; see Appendix D.5 for a more detailed discussion.

The last element we need for the statement of our results is the minimum total cost of reaching a component $\mathcal{K}_i$ of crit f from any starting point. Since the most likely trajectories of (SGD) are action minimizers, the "path of least resistance" to reach vertex i from vertex j may not follow the edge (i, j) if the cost B_{ij} is too high; instead, the relevant notion turns out to be the *minimum weight spanning tree* pointing to i.[7] Formally, writing $\mathcal{T}_i$ for the set of spanning trees of $\mathcal{G}$ that point to i, we define the *energy* of $\mathcal{K}_i$ as

$$E_i = \min_{T_i \in \mathcal{T}_i} \sum_{j,k \in T_i} B_{jk}. \tag{21}$$

The terminology "energy" is explained below, where we show that, to leading order, the long-run distribution of (SGD) around crit f follows the Boltzmann–Gibbs distribution for a canonical ensemble with energy levels E_i at temperature η.

3.4. **The long-run distribution of SGD.** With all this in hand, we are finally in a position to state our main results for the statistics of the asymptotic behavior of (SGD). We start by showing that, in the long run, the probability of observing the iterates of (SGD) near a component of crit f is exponentially proportional to its energy.

Theorem 1. *Suppose that μ_∞ is invariant under (SGD), fix a tolerance level $\varepsilon > 0$, and let $\mathcal{U}_i \equiv \mathcal{U}_i(\delta)$, $i = 1, \dots, K$, be δ-neighborhoods of the components of* crit f. *Then, for all sufficiently small $\delta, \eta > 0$, we have*

$$|\eta \log \mu_\infty(\mathcal{U}_i) + E_i - \min_j E_j| \leq \varepsilon \tag{22}$$

and

$$\left| \eta \log \frac{\mu_\infty(\mathcal{U}_i)}{\mu_\infty(\mathcal{U}_j)} + E_i - E_j \right| \leq \varepsilon. \tag{23}$$

[7] We distinguish here between the notion of an *out-tree* and that of an *in-tree*. In an out-tree, edges point away from the root; in an in-tree, edges point toward it [62].

More compactly, with notation as above, we have:

$$\mu_\infty(\mathcal{U}_i) \propto \exp\left(-\frac{E_i + \mathcal{O}(\varepsilon)}{\eta}\right). \tag{24}$$

Theorem 1 is the formal version of the statement that the long-run distribution of (SGD) around the components of crit f follows an ε-approximate Boltzmann–Gibbs distribution with energy levels E_i at temperature η [38]. However, since the critical set of f includes both minimizing and *non-minimizing* components,[8] a natural question that arises is whether the non-minimizing components of f are selected against under μ_∞. Our next result is a consequence of Theorem 1 and shows that this indeed the case:

Theorem 2. *Suppose that μ_∞ is invariant under (SGD), and let $\mathcal{K}$ be a non-minimizing component of f. Then, with notation as in Theorem 1, there exists a minimizing component $\mathcal{K}'$ of f and a positive constant $c \equiv c(\mathcal{K}, \mathcal{K}') > 0$ such that*

$$\frac{\mu_\infty(\mathcal{U})}{\mu_\infty(\mathcal{U}')} \leq \exp\left(-\frac{c(\mathcal{K}, \mathcal{K}') + \varepsilon}{\eta}\right) \tag{25}$$

for all all sufficiently small $\eta > 0$ and all sufficiently small neighborhoods $\mathcal{U}$ and $\mathcal{U}'$ of $\mathcal{K}$ and $\mathcal{K}'$ respectively. In particular, in the limit $\eta \to 0$, we have $\mu_\infty(\mathcal{U}) \to 0$.

This avoidance principle is particularly important because it shows that (SGD) is far less likely to be observed near a non-minimizing components of crit f relative to a minimizing one. In this regard, Theorem 2 complements a broad range of avoidance results in the literature [21, 26, 27, 31, 52, 58] without requiring any of the "strict saddle" assumptions that are standard in this context.

That being said, Theorems 1 and 2 leave open the possibility that the energy landscape of (SGD) contains *non-critical* low-energy regions that nonetheless get a significant amount of probability under (SGD); put differently, Theorems 1 and 2 do not rule out the eventuality that, in the long run, x_n may still be observed with non-vanishing probability far from the critical region of f. Our next result addresses precisely this issue and shows that this probability is exponentially small.

Theorem 3. *Suppose that μ_∞ is invariant under (SGD), and let $\mathcal{U} \equiv \mathcal{U}(\delta)$ be a δ-neighborhood of* crit f*. Then there exists a constant $c \equiv c_\delta > 0$ such that, for all sufficiently small $\eta > 0$, we have:*

$$\mu_\infty(\mathcal{U}) \geq 1 - e^{-c/\eta}. \tag{26}$$

Taken together, Theorems 1–3 show that, in the long run, the iterates of (SGD) are exponentially more likely to be observed in the vicinity of crit f rather than far from it, and exponentially more likely to be observed near a minimum of f rather than a saddle-point (or a local maximizer).

Our next result can be seen as joint consequence of Theorems 1 and 3 as it shows that the long-run distribution of (SGD) is exponentially concentrated around the system's *ground state*

$$\mathcal{K}_0 = \bigcup_{i \in \arg\min_j E_j} \mathcal{K}_i \tag{27}$$

that is, the components of crit f with minimal energy. The precise statement is as follows:

[8] To remove any ambiguity, a component $\mathcal{K}$ of crit f is called (locally) minimizing if $\mathcal{K} = \arg\min_{x \in \mathcal{U}} f(x)$ for some neighborhood $\mathcal{U}$ of $\mathcal{K}$; otherwise, we say that $\mathcal{K}$ is non-minimizing.

Theorem 4. *Suppose that μ_∞ is invariant under (SGD), and let $\mathcal{U}_0 \equiv \mathcal{U}_0(\delta)$ be a δ-neighborhood of the system's ground state $\mathcal{K}_0$. Then there exists a constant $c \equiv c_\delta > 0$ such that, for all sufficiently small $\eta > 0$, we have:*

$$\mu_\infty(\mathcal{U}_0) \geq 1 - e^{-c/\eta}. \tag{28}$$

In words, Theorems 1–4 provide the following quantification of the limiting distribution of x_n: in the long run (a) the critical region of f is visited exponentially more often than any non-critical region of f (Theorem 3); (b) in particular, the iterates of (SGD) are exponentially concentrated around the problem's ground state (Theorem 4); (c) among the mass that remains, every component of f gets a fraction that is exponentially proportional to its energy (Theorem 1); and, finally (d) every non-minimizing component is "dominated" by a minimizing component that is visited exponentially more often (Theorem 2). Importantly, the problem's energy landscape is shaped by f, but not f alone: the statistics of (SFO) play an equally important role, so we may have $\mathcal{K}_0 \neq \arg\min f$; we discuss this issue in detail in Section 4.

The proofs of the above results are quite lengthy and elaborate, so we defer them to the appendix and only provide below a roadmap describing the overall proof strategy, the main technical challenges encountered, and the way they can be resolved.

3.5. **Outline of the proof.** As discussed in Section 3.2, the first step of the proof consists in establishing a large deviation principle for (SGD). The LDP of Proposition 1 for the interpolated process $X(t)$ is obtained as a consequence of [20, Chap. 7] in Appendix C.2. With this result in hand, our next step (which we carry out in Appendix C.3) is to deduce an LDP for the "accelerated" SGD process x_n^η, $n = 0, 1, \dots$, defined here as

$$x_n^\eta = x_{n\lfloor 1/\eta \rfloor} \tag{29}$$

where $\lfloor 1/\eta \rfloor$ denotes the integer part of $1/\eta$. As x_n^η is a subsampled version of (SGD), it essentially shares the same long-run behavior and invariant measures and, in addition, it has a very important feature: there are sufficiently many time steps between two of its iterates for an LDP to hold in the specified subinterval. [Intuitively, this is because it takes $\mathcal{O}(1/\eta)$ steps of (SGD) to average out random fluctuations due to the noise.] In view of the above, the rest of our proof focuses on the accelerated sequence x_n^η.

The main thrust of the proof is contained in Appendix D and consists in adapting the powerful machinery of [20, 35] to study the limiting behavior of x_n^η. However, both [20, 35] study continuous-time diffusion processes on closed manifolds, so there are some key challenges to overcome:

- The unconstrained setting renders many elements of [20, 35] inapplicable. We remedy this by showing that the time that (SGD) only spends a negligible amount of time away from $\operatorname{crit} f$.
- The generality of our assumptions on the noise makes the Lagrangians $\mathcal{L}_\mathsf{G}$ and $\mathcal{L}_\mathsf{Z}$ non-smooth: more precisely, they must have bounded domains, on which they may fail to be continuous. As a consequence, most of the objects introduced in [20, 35] become drastically less regular – e.g., B defined in (18) – which again renders their results inapplicable. We remedy this issue by refining the analysis, carefully studying the structure of the attractors, and salvaging enough regularity to proceed.

The crux of the proof (Appendix D) is structured as follows:

(1) In Appendix D.2, we study the structure of the attractors of (SGD), as well as the regularity of the Lagrangians and the quasi-potential near these attractors.

(2) In Appendix D.3, we show that (SGD) spends most of its time near its attractors by deriving a series of tail-bounds on the time spent away from crit f. These bounds are obtained even for unbounded variance proxy through the construction of a Lyapunov function from f and σ_∞^2.
(3) In Appendix D.5, we estimate the transition probabilities of the process between attractors: if the iterates of (SGD) are near $\mathcal{K}_i$, the next component of critical points that they visit is $\mathcal{K}_j$ with probability $\exp(-B_{ij}/\eta)$. As such, low-weight paths in the transition graph $\mathcal{G}$ of (SGD) represent the high-probability transitions of (SGD) between components. We can then leverage the general theory of Freidlin & Wentzell [20] to obtain Theorem 1.
(4) Finally, in Appendix D.6, we analyze the properties of minimizing components to establish Theorems 2–4 .

Remark. We also note that, since weak limit points of the sequence of occupation measures μ_n of (SGD) are invariant in the sense of (9), the results of Theorems 1–4 also apply to μ_n if n is large enough. We make this observation precise in Appendix D.8. ◀

4. Examples and applications

In this last section, we explore the dependency of the transition costs and the energy levels on the parameters of the problem in certain special cases.

4.1. Gaussian noise.

We begin with the case of (truncated) Gaussian noise, where the problem's energy levels admit a particularly simple closed form. To ease notation, we present here the computations in the case of Gaussian noise, and we defer the more intricate case of *truncated* Gaussian noise to Appendix E.2.

To that end, assume that the gradient error $Z(x,\omega)$ in (SGD) follows a centered Gaussian distribution with variance $\sigma^2 > 0$ for all $x \in \mathcal{X}$. The Lagrangian and the action functional of the problem then become:

$$\mathcal{L}_{\mathsf{G}}(x,v) = \frac{\|v + \nabla f(x)\|^2}{2\sigma^2} \tag{30}$$

and

$$\mathcal{S}_T[\gamma] = \int_0^T \frac{\|\dot\gamma(t) + \nabla f(\gamma(t))\|^2}{2\sigma^2}\, dt \tag{31}$$

for all $x, v \in \mathcal{X}$ and all $\gamma \in \mathcal{C}_T$. This expression shows that $\mathcal{S}_T[\gamma]$ penalizes the deviation of γ from the gradient flow of f: the closer $\dot\gamma(t)$ is to $-\nabla f(\gamma(t))$, the smaller the action. Then, for the reverse path $\varphi(t) = \gamma(T-t)$, we get

$$\mathcal{S}_T[\varphi] = \mathcal{S}_T[\gamma] - 2[f(\gamma_T) - f(\gamma_0)]/\sigma^2. \tag{32}$$

Note that if γ joins $\mathcal{K}_i$ to $\mathcal{K}_j$, then φ joins $\mathcal{K}_j$, so

$$B_{ji} \leq B_{ij} - 2(f_j - f_i)/\sigma^2 \tag{33}$$

where f_i denotes the value of f on $\mathcal{K}_i$, $i = 1, \dots, K$. Exchanging i and j, we get equality in (33), so the minimum in the definition (21) of E_i is reached for the same undirected tree, namely the minimum-weight spanning tree with symmetric weights $B_{ij} + 2f_i/\sigma^2$ for all i, j. As such, up to a constant, we get

$$E_i = 2f_i/\sigma^2. \tag{34}$$

Thus, invoking Theorem 1, we conclude that the probabiity distribution of (SGD) over $\operatorname{crit}(f)$ is governed by the Boltzmann–Gibbs measure with energy levels given by (34)

Similarly, in the *truncated* Gaussian case, we have:

Proposition 2. *For any $\varepsilon > 0$, if $Z(x,\omega)$ follows a centered Gaussian distribution with variance $\sigma^2 > 0$ conditioned on being in a ball with large enough radius depending on $\nabla f(x)$ and ε, then, up to a constant,*

$$E_i = 2f_i/\sigma^2 + \mathcal{O}(\varepsilon) \quad \text{for all } i = 1,\dots,K\,. \tag{35}$$

Proposition 2 is proven in Appendix E.2, where we also allow σ to depend on x via $f(x)$.

4.2. **Local dependencies.** The modeling of the noise is crucial to the understanding of the dynamics of (SGD). This has been often underlined, especially in the exit-time literature [28, 30, 56, 68]. In particular, [56] showed experimentally that, in deep learning models, the variance of the noise scales linearly with the objective function and also examined the effect of this observation on the exit time from a local minima.

We explain here how this model of the noise influences the invariant measure of (SGD). To that end, following Mori et al. [56], assume there is a positive-definite matrix H^* such that, locally near a minimizing $\mathcal{K}_i$, $\log f$ is separable in the eigenbasis of H^*:[9]

$$\log f(x) = \sum_{\lambda\in\operatorname{eig} H^*} g^\lambda(x_\lambda) \tag{36}$$

where x_λ denotes the projection of x on the eigenspace of λ.

Lemma 1. *Suppose that $Z(x,\omega)$ satisfies*

$$K_{\mathsf{Z}}(x,v) \leq \frac{\sigma^2 f(x)}{2}\langle v, H^* v\rangle \quad \text{for all } x \text{ near } \mathcal{K}_i\,. \tag{37}$$

Then, for small enough $\delta > 0$ and all $j \neq i$, we have

$$E_j \geq 2\min\left\{\sum_{\lambda\in\operatorname{eig} H^*} \frac{g^\lambda(x_\lambda) - g_i^\lambda}{\lambda\sigma^2} : x, \operatorname{dist}(x,\mathcal{K}_i) = \delta\right\} \tag{38}$$

where g_i^λ is the value of g^λ on $\mathcal{K}_i$.

This result shows that the energy of each component of $\operatorname{crit}(f)$ is lower bounded by the RHS of (38). This quantity scales as the reciprocal of the eigenvalues of H^* so, as the minimum becomes flatter (i.e., the eigenvalues of H^* become smaller), the energy levels of all other components become larger: thus, relative to component $\mathcal{K}_i$, the other components all become less probable. Moreover, note that the RHS of (38) only scales logarithmically with the value of the objective function around $\mathcal{K}_i$, i.e., the depth of the minimum: this means that the "flatness" of the minimum plays a greater role in the relative probabilities of the components than the depth.

This also shows that deepest minima do not necessarily correspond to the ground state of the problem: if σ^2 or the eigenvalues of H^* is small enough compared to the noise level outside the δ-neighborhood of $\mathcal{K}_i$, then $\mathcal{K}_i$ will be the ground state of the system even if it is not the deepest minimum; we provide a formal proof of this in Appendix E.3.

5. Concluding remarks

Our objective was to quantify the long-run distribution of SGD in a general, non-convex setting. As far as we are aware, our paper provides the first description of the invariant measure of (SGD), and in particular, its distribution over components of critical points. This distribution is governed by energy levels that depend both on the objective function and the statistics of the noise. An important challenge that remains is to estimate these energy

[9] Following [56], H^* corresponds to the Hessian of f at $\mathcal{K}_i$ for deep learning models.

levels in different settings; this would be a major step towards a better understanding of the generalization properties of SGD.

Acknowledgments

This research was supported in part by the French National Research Agency (ANR) in the framework of the PEPR IA FOUNDRY project (ANR-23-PEIA-0003), the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003). PM is also a member of the Archimedes Research Unit/Athena RC, and was partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

Appendix A. Further related work

A.1. **Consequences of the diffusion approximation.** The SDE approximation of SGD, introduced by Li et al. [42, 43], has been a fruitful development in the understanding of some aspects of the dynamics of SGD. For instance, Ziyin et al. [71] provide explicit descriptions for the invariant measure of the diffusion approximation of SGD for diagonal linear neural networks. Applications of this SDE approximation also include the study the dynamics of SGD close to manifold of minimizers [6, 45]. Wojtowytsch [67] study the invariant measure of the diffusion approximation: if the set of global minimizers form a manifold on which the noise vanishes, they show that the invariant measure of the diffusion concentrates on this manifold and moreover provide a description of the limiting measure on this manifold.

Another line of works focuses on the case where the objective function is scale-invariant [44] and how this impacts the convergence of the dynamics of SGD: Wang & Wang [66] describes the convergence of the SDE approximation with anisotropic constant noise to the Gibbs measure, while Li et al. [46] shows that discrete-time SGD dynamics close to a manifold of minimizers enjoy fast convergence to an invariant measure.

Finally, Mignacco & Urbani [54], Mignacco et al. [55], Veiga et al. [64] leverage dynamic mean-field theory (DMFT) to study the behavior of the diffusion approximation of SGD. The DMFT, or "path-integral" approach, comes from statistical physics and bears a close resemblance to the Freidlin-Wentzell theory of large deviations for SDEs. However, this methodology, as well as Mignacco & Urbani [54], Mignacco et al. [55], Veiga et al. [64], is restricted to the continuous-time diffusion and remains at heuristic level.

Let us underline two points comparing these works to ours. While these works focus on the learning behavior of the algorithm by considering specific statistical models and specific losses, we focus on the optimization aspects and on covering general non-convex objectives. Secondly, these results are either local or concern the asymptotic distribution of the continuous-time approximation of (SGD), which do no provide information of the asymptotic behaviour of the actual discrete-time dynamics.

A.2. **On the heavy-tail character of the asymptotic distribution of SGD.** A recent line of work has focused on the heavy-tail character of the asymptotic distribution of (SGD) [23, 25, 57]. These works show that, under some broad conditions, the stationary distribution of (SGD) is heavy-tailed: specifically, for some $\alpha > 0$, the tails of the stationary distribution x_∞ of the iterates of (SGD) decays as $\mathbb{P}(\|x_\infty\| \geq z) = \Theta(z^{-\alpha})$ or $\mathbb{P}(u^\top x_\infty \geq z) = \Theta(z^{-\alpha})$ for all $u \in \mathbb{R}^d$. As such, these results concern the probability of observing the iterates of SGD at very large distances from the origin. This is in contrast with our work, which focuses on the distribution of the iterates of SGD near critical regions of the objective function. These two types of results are thus orthogonal and complementary.

A.3. **SGD with a vanishing step-size.** The long-run behavior of (SGD) is markedly different when the method is run with a vanishing step-size $\eta_n > 0$ with $\lim_{n\to\infty} \eta_n = 0$. This was the original version of (SGD) as proposed by Robbins & Monro [61] – and, in a slightly modified form by Kiefer & Wolfowitz [34] – and, in contrast to the constant step-size case, the vanishing step-size algorithm converges with probability 1. The first almost sure convergence result of this type was obtained by Ljung [48] under the assumption that the iterates of (SGD) remain bounded. This boundedness assumption was dropped by Benaïm [3] and Bertsekas & Tsitsiklis [5] who showed that (SGD) with a vanishing step-size converges to a component of crit f as long as η_n satisfies the Robbins–Monro summability conditions $\sum_n \eta_n = \infty$ and $\sum_n \eta_n^2 < \infty$; for a series of related results under different assumptions, cf. [52, 53, 70] and references therein. In all these cases, ergodicity of the process is lost (because of the vanishing step-size), so the limiting distribution of (SGD) depends crucially on its initialization and other non-tail events. We are not aware of any results quantifying the long-run distribution of (SGD) with a vanishing step-size.

Appendix B. Setup and Preliminaries

B.1. **Notation.** Throughout the sequel, we will write $\langle \cdot, \cdot \rangle$ for the standard inner product on $\mathcal{X} \equiv \mathbb{R}^d$ and $\|\cdot\|$ for the induced (Euclidean) norm. To lighten notation, we will identify $\mathcal{X}$ with its dual $\mathbb{R}^d \equiv \mathcal{X}^*$, and we will not formally distinguish between primal and dual vectors (though the distinction should be clear from the context). We will also write $\mathbb{B}(x, r)$ (resp. $\overline{\mathbb{B}}(x, r)$) for the open (resp. closed) ball of radius r centered at $x \in \mathcal{X}$, and we will respectively denote the open and closed δ-neighborhoods of $S \subseteq \in \mathcal{X}$ as

$$\mathcal{U}_\delta(S) \coloneqq \{x \in \mathcal{X} : \operatorname{dist}(S, x) < \delta\} \tag{B.1.1a}$$

$$S_\delta \coloneqq \operatorname{cl}(\mathcal{U}_\delta(S)) = \{x \in \mathcal{X} : \operatorname{dist}(S, x) \leq \delta\} \tag{B.1.1b}$$

B.2. **Setup and assumptions.** Before we begin our proof, we revisit and discuss our standing assumptions. In particular, to extend the range of our results, we provide in the rest of this appendix a weaker version of the blanket assumptions of Section 2, which we label with an asterisk ("$*$") and which will be in force throughout the appendix.

We begin with our assumptions for the objective function f of (Opt).

Assumption 1* (Weaker version of Assumption 1)**.** The objective function $f\colon \mathcal{X} \to \mathbb{R}$ satisfies the following conditions:

(*a*) *Coercivity:* $f(x) \to \infty$ as $\|x\| \to \infty$.

(*b*) *Smoothness:* f is C^2-differentiable and its gradient is β-Lipschitz continuous, that is,

$$\|\nabla f(x') - \nabla f(x)\| \leq \beta \|x' - x\| \quad \text{for all } x, x' \in \mathcal{X}\,. \tag{LG}$$

(*c*) *Critical set regularity:* The critical set

$$\operatorname{crit} f \coloneqq \{x \in \mathcal{X} : \nabla f(x) = 0\} \tag{1}$$

of f consists of a finite number of essentially smoothly connected components $\mathcal{K}_i$, $i = 1, \dots, K$.

The difference between Assumptions 1 and 1* is that, in the latter, the connected components of crit f are only required to be *essentially* smoothly connected. Formally, this means that, for any connected component $\mathcal{K}$ of crit f, and for any two points $x, x' \in \mathcal{K}$, there exists (*a*) a continuous, almost everywhere differentiable curve $\gamma\colon [0,1] \to \mathcal{K}$ such that $\gamma(0) = x$, $\gamma(1) = x'$; and (*b*) a partition $0 = t_0 < \cdots < t_n < \cdots < t_N = 1$ of $[0,1]$ such that γ is integrable on every closed interval of (t_{n-1}, t_n) for all $n = 1, \dots, N$.

Remark B.1. The path-connectedness requirement of Assumption 1(c) is satisfied whenever the connected components of crit f are isolated critical points, smooth manifolds, or finite unions of closed manifolds. More generally, Assumption 1*(c) is satisfied whenever f is definable – in which case crit f is also definable, so each component can be connected by piecewise smooth paths [9, 63]. The relaxation provided by Assumption 1*(c) represents the "minimal" set of hypotheses that are required for our analysis to go through. ◀

Moving forward, to align our notation with standard conventions in large deviations theory, it will be more convenient to work with $-\mathsf{Z}(x;\omega)$ instead of $\mathsf{Z}(x;\omega)$ in our proofs. To make this clear, we restate below Assumption 2 in terms of the noise process

$$\mathsf{U}(x,\omega) = -\mathsf{Z}(x;\omega) \tag{B.2.1}$$

and we make use of this opportunity to relax the definition of the variance proxy of U.

Assumption 2* (Weaker version of Assumption 2)**.** The noise term $\mathsf{U}\colon \mathcal{X}\times\Omega\to\mathbb{R}^d$ satisfies the following properties:

(a) *Properness:* $\mathbb{E}[\mathsf{U}(x,\omega)] = 0$ and $\operatorname{cov}(\mathsf{U}(x,\omega)) \succ 0$ for all $x\in\mathcal{X}$.

(b) *Smooth growth:* $\mathsf{U}(x,\omega)$ is C^2-differentiable and satisfies the growth condition

$$\sup_{x,\omega} \frac{\|\mathsf{U}(x,\omega)\|}{1+\|x\|} < \infty\,. \tag{2}$$

(c) *Sub-Gaussian tails:* The cumulant-generating function of U is bounded as

$$\log \mathbb{E}\Big[e^{\langle p,\mathsf{U}(x,\omega)\rangle}\Big] \leq \frac{1}{2}\sigma_\infty^2(f(x))\,\|p\|^2 \quad \text{for all } p\in\mathbb{R}^d, \tag{3*}$$

where $\sigma\colon\mathbb{R}\to(0,\infty)$ is continuous, bounded away from zero ($\inf\sigma_\infty^2 > 0$), and grows at infinity as $\sigma_\infty^2(f(x)) = \Theta(\|x\|^s)$ for some $s\in[0,2]$, i.e.,

$$0 < \liminf_{\|x\|\to\infty} \frac{\sigma_\infty^2(f(x))}{\|x\|^s} \leq \limsup_{\|x\|\to\infty} \frac{\sigma_\infty^2(f(x))}{\|x\|^s} < \infty\,. \tag{B.2.2}$$

Remark B.2. The difference between Assumptions 2 and 2* lies in the requirements for the variance proxy σ_∞^2 of the noise in (SGD). Because f is coercive, the dependence of σ_∞^2 on $f(x)$ allows for noise processes with an unbounded variance as $\|x\|\to\infty$; the specific functional dependence through $f(x)$ instead of x is a modeling choice which we make because it greatly simplifies the proof and our calculations. ◀

In this more general setting, we augment Assumption 3 with a technical condition that is satisfied trivially when σ_∞^2 is constant.

Assumption 3* (Weaker version of Assumption 3)**.** The signal-to-noise ratio of G is bounded as

$$\liminf_{\|x\|\to\infty} \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} > 16\log 6\cdot d\,. \tag{4*}$$

► **Example B.1** (Example of Section 2.2, redux)**.** In Section 2.2, we discussed the regularized empirical risk minimization problem and mentioned that, under a dissipativity condition and with a large enough batch size, it fits our framework. We now provide more details.

Consider the objective f given by

$$f(x) = \frac{1}{n}\sum_{i=1}^n \ell(x;\xi_i) + \frac{\lambda}{2}\|x\|^2 \tag{B.2.3}$$

where $\ell(x;\xi)$ represents the loss of the model x on the data point ξ, ξ_i, $i = 1,\dots,n$, are the training data and λ is the (positive) regularization parameter.

Let us assume that ℓ is non-negative, C^2-differentiable, β-Lipschitz smooth and that the resulting objective f is dissipative: there are $\alpha, \beta > 0$ such that

$$\langle \nabla f(x), x\rangle \geq \alpha\|x\|^2 - \beta\,. \tag{B.2.4}$$

As is usually the case in practice, we consider the SFO obtained by sampling mini-batches of size B. The noise term $\mathsf{U}(x,\omega)$ is then given by

$$\mathsf{U}(x,\omega) = \frac{1}{B}\sum_{b=1}^{B} \nabla\ell(x;\xi_{\omega_b}) - \frac{1}{n}\sum_{i=1}^{n}\nabla\ell(x;\xi_i) \tag{B.2.5}$$

with $\omega = (\omega_1,\dots,\omega_B)$ representing B indices from $\{1,\dots,n\}$.

All the terms

$$\nabla\ell(x;\xi_{\omega_b}) - \frac{1}{n}\sum_{i=1}^{n}\nabla\ell(x;\xi_i) \tag{B.2.6}$$

are uniformly bounded by $\mathcal{O}(\|x\|)$ by smoothness of ℓ. In particular, this implies that Assumption 2*(*b*) is satisfied.

Moreover, we also obtain that all the terms of the form Eq. (B.2.6) are $\mathcal{O}\big(\|x\|^2\big)$-sub-Gaussian, and therefore, by independence we obtain that $\mathsf{U}(x,\omega)$ is $\mathcal{O}\big(\frac{1}{B}\|x\|^2\big)$-sub-Gaussian.

Since ℓ is non-negative, f is lower-bounded by $\Omega(\|x\|^2)$. It gives that $\mathsf{U}(x,\omega)$ is actually $\mathcal{O}\big(\frac{1}{B}f(x)\big)$-sub-Gaussian and therefore Assumption 2*(*b*) is satisfied with $\sigma_\infty^2(t) \propto \frac{t}{B}$.

We now show that Assumption 3* can be satisfied by choosing B large enough. Indeed, by dissipativity we have that

$$\begin{aligned}\|\nabla f(x)\|^2 &= \|\nabla f(x) - \alpha x\|^2 + \alpha^2\|x\|^2 + 2\alpha\langle \nabla f(x) - \alpha x, x\rangle \\ &\geq \alpha^2\|x\|^2 - 2\beta = \Omega(\|x\|^2)\,,\end{aligned} \tag{B.2.7}$$

so that

$$\frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \geq \Omega\left(\frac{\|x\|^2}{f(x)/B}\right) = \Omega(B)\,. \tag{B.2.8}$$

The second part of Assumption 3* is satisfied by non-negativity and smoothness of ℓ. ◀

In this framework, the iterates of (SGD), started at $x \in \mathcal{X}$, are defined by the following recursion:

$$\begin{cases} x_0 \in \mathcal{X} \\ x_{n+1} = x_n - \eta\nabla f(x_n) + \eta\mathsf{U}_n\,, \quad \text{where } \mathsf{U}_n = \mathsf{U}(x_n,\omega_n) \end{cases} \tag{B.2.9}$$

where $(\omega_n)_{n\geq 0}$ is a sequence of random variables in $\mathbb{R}^m$. We will denote by $\mathbb{P}_x$ the law of the sequence $(\omega_n)_{n\geq 0}$ when the initial point is x and by $\mathbb{E}_x$ the expectation with respect to $\mathbb{P}_x$.

Assumptions 1* and 2* imply the following growth condition, that we assume holds with the same constant for the sake of simplicity. There is $M > 0$ such that, for all $x \in \mathcal{X}$, $\omega \in \Omega$,

$$\|\nabla f(x)\| \leq M(1+\|x\|) \quad \text{and} \quad \|\mathsf{U}(x,\omega)\| \leq M(1+\|x\|)\,. \tag{B.2.10}$$

We introduce the cumulant generating functions of the noise $\mathsf{U}(x,\omega)$ and of the drift $-\nabla f(x) + \mathsf{U}(x,\omega)$, that we denote by $\bar{\mathcal{H}}$, $\mathcal{H}$ to avoid confusion. We also define their convex conjugates, that we denote by $\bar{\mathcal{L}}$, $\mathcal{L}$.

Definition 1 (Hamiltonians and Lagrangians)**.** Define, for $x \in \mathcal{X}$, $v \in \mathbb{R}^d$,

$$\begin{aligned} \bar{\mathcal{H}}(x,v) &= \log \mathbb{E}[\exp(\langle v, \mathsf{U}(x,\omega)\rangle)] \\ \mathcal{H}(x,v) &= -\langle \nabla f(x), v\rangle + \bar{\mathcal{H}}(x,v) \\ \bar{\mathcal{L}}(x,v) &= \bar{\mathcal{H}}(x,\cdot)^*(v) \\ \mathcal{L}(x,v) &= \mathcal{H}(x,\cdot)^*(v) = \bar{\mathcal{L}}(x, v + \nabla f(x))\,. \end{aligned} \tag{B.2.11}$$

$\bar{\mathcal{L}}$ and $\mathcal{L}$ are thus respectively equal to the Lagrangians $\mathcal{L}_{\mathsf{Z}}(\cdot,\cdot)$ and $\mathcal{L}_{\mathsf{G}}(\cdot,\cdot)$. Finally, Assumption 4 will be discussed in Appendix D.5.

B.3. **Basic properties.** In this section, we derive from our assumptions some basic consequences, which will be useful throughout the proof.

We first state some properties of the Hamiltonian and the Lagrangian, which follow from their definitions.

Lemma B.1 (Properties of $\mathcal{H}$ and $\mathcal{L}$)**.**

(1) $\mathcal{H}$ is $\mathcal{C}^2$ and $\mathcal{H}(x,\cdot)$ is convex for any $x \in \mathcal{X}$.

(2) $\mathcal{L}(x,\cdot)$ is convex for any $x \in \mathcal{X}$, $\mathcal{L}$ is lower semi-continuous (l.s.c.) on $\mathcal{X} \times \mathbb{R}^d$.

(3) For any $x \in \mathcal{X}$, $v \in \mathbb{R}^d$, $\mathcal{H}(x,v) \le 2M(1+\|x\|)\|v\|^2$ and $\operatorname{dom} \mathcal{L}(x,\cdot) \subset \overline{\mathbb{B}}(0, 2M(1+\|x\|))$.

(4) For any $x \in \mathcal{X}$, $v \in \mathbb{R}^d$, $\mathcal{L}(x,v) \ge 0$ and $\mathcal{L}(x,v) = 0 \iff v = \nabla f(x)$.

Proof. For the last point, since $\mathcal{H}(x,\cdot)$ is convex and differentiable,

$$\mathcal{L}(x,v) = 0 \iff 0 \in \partial_v \mathcal{L}(x,v) \iff v \in \partial_v \mathcal{H}(x,0) \iff v = \nabla_v \mathcal{H}(x,0)\,, \tag{B.3.1}$$

and, since the noise has zero mean, $\nabla_v \mathcal{H}(x,0) = \nabla f(x)$. ∎

Lemma B.2 (Growth of the iterates)**.** *For any $x_0 \in \mathcal{X}$, $\eta_0 > 0$, for every $\eta \in (0, \eta_0]$, for any $N \ge 1$, $0 \le n \le N\lceil \eta^{-1} \rceil$,*

$$\|x_n\| \le e^{2M(1+\eta_0)N}(1 + \|x_0\|)\,. \tag{B.3.2}$$

Proof. By the triangular inequality, for any $n \ge 0$, we have that

$$\begin{aligned} \|x_{n+1}\| &\le \|x_n\| + \eta\|\nabla f(x_n)\| + \eta\|\mathsf{U}_n\| \\ &\le \|x_n\| + 2\eta M(1 + \|x_n\|) \\ &= (1 + 2\eta M)\|x_n\| + 2\eta M \end{aligned} \tag{B.3.3}$$

where we used the growth condition on ∇f and U. One can then solve this recursion to show that, for any $n \ge 0$,

$$\|x_n\| \le (1 + 2\eta M)^n (1 + \|x_0\|)\,. \tag{B.3.4}$$

The result then follows by noticing that, for $0 \le n \le N\lceil \eta^{-1} \rceil$,

$$\begin{aligned} (1 + 2\eta M)^n &\le (1 + 2\eta M)^{N\lceil \eta^{-1} \rceil} \\ &\le e^{2MN\eta\lceil \eta^{-1} \rceil} \\ &\le e^{2MN(1+\eta)}\,. \end{aligned}$$

∎

Lemma B.3. *For any $x_0 \in \mathcal{X}$, for every $\eta \in (0, (4M)^{-1}]$, for any $N \ge 1$, $0 \le n \le N\lceil \eta^{-1} \rceil$,*

$$\|x_n\| \ge e^{-(4M+1)N}\|x_0\| - 1\,. \tag{B.3.5}$$

Proof. By the triangular inequality, for any $n \geq 0$, we have that

$$\begin{aligned}\|x_{n+1}\| &\geq \|x_n\| - \eta\|\nabla f(x_n)\| - \eta\|\mathsf{U}_n\| \\ &\leq \|x_n\| - 2\eta M(1 + \|x_n\|) \\ &= (1 - 2\eta M)\|x_n\| - 2\eta M \end{aligned} \tag{B.3.6}$$

where we used the growth condition on ∇f and U. One can then solve this recursion to obtain that, for any $n \geq 0$,

$$\|x_n\| \geq (1 - 2\eta M)^n \|x_0\| - 1\,. \tag{B.3.7}$$

The result then follows by noticing that, for $0 \leq n \leq N\lceil \eta^{-1} \rceil$,

$$\begin{aligned}(1 - 2\eta M)^n &\geq (1 - 2\eta M)^{N\lceil \eta^{-1} \rceil} \\ &\geq e^{-4MN\eta\lceil \eta^{-1} \rceil} \\ &\leq e^{-4MN(1+\eta)} \end{aligned} \tag{B.3.8}$$

where we used the fact that $\log(1 - x) \geq -2x$ on $[0, 1/2]$ since $2\eta M \leq 1/2$. ∎

Appendix C. A large deviation principle for SGD

C.1. **Preliminaries.** The goal of this section is to provide large deviation principles for continuous and discrete trajectories related to our algorithm of interest (SGD). From the sequences $(x_n)_{n\geq 0}$ and $(\omega_n)_{n\geq 0}$, we define three sequences: one discrete (in lowercase) and two continuous (in uppercase):

(a) The discrete "rescaled" trajectory

$$x_n^\eta := x_{n\lfloor 1/\eta \rfloor}\,. \tag{C.1.1a}$$

(b) The continuous "interpolated" trajectory, defined for any $n \geq 0$, $t \in [\eta n, \eta(n+1)]$ by

$$X_t = x_n + \left(\frac{t}{\eta} - n\right)(x_{n+1} - x_n) \tag{C.1.1b}$$

(c) The continuous "discretized noise" trajectory, defined by $Z_0 = x_0$ and for any $t > 0$

$$\dot{Z}_t = -\nabla f(Z_t) + \mathsf{U}(Z_t, \omega_{\lfloor t/\eta \rfloor}) \tag{C.1.1c}$$

Note that all three sequences are "accelerated" by a factor $1/\eta$ compared to the original sequences appearing in (SGD). In this section, we establish a large deviation principle for the discrete rescaled sequence $(x_n^\eta)_{n\geq 0}$. To do so, we build upon Freidlin & Wentzell [20, Chap. 7] to obtain a large deviations principle on $(Z_t)_{t\geq 0}$ and then transfer it to $(X_t)_{t\geq 0}$ which enables us to obtain a discrete large deviation principle for $(x_n^\eta)_{n\geq 0}$.

We equip the space of continuous functions $\mathcal{C}_T := \mathcal{C}([0,T], \mathcal{X})$ with the distance induced by the uniform norm

$$\mathrm{dist}_{[0,T]}(\gamma, \varphi) := \sup_{t\in[0,T]} \|\gamma_t - \varphi_t\|\,. \tag{C.1.2}$$

In order to use it as a proxy later, we will now bound the distance between the (continuous) "interpolated" trajectory and the "discretized noise" trajectory. To do so, we will first bound the latter.

Lemma C.1 (Growth of the trajectory)**.** *For any $x_0 \in \mathcal{X}$, $\eta > 0$, $t \geq 0$, we have $\|Z_t\| \leq e^{2Mt}(\|x_0\| + 2Mt)$.*

Proof. Using the definition of Z_t in (C.1.1c) and the growth condition (B.2.10) on ∇f and U, we have that

$$\|\dot{Z}_t\| = \left\|-\nabla f(Z_t) + \mathsf{U}(Z_t, \omega_{\lfloor t/\eta \rfloor})\right\| \leq 2M(1 + \|Z_t\|)\,. \tag{C.1.3}$$

Hence, for any $t \geq 0$,

$$\|Z_t\| \leq \|z_0\| + \int_0^t 2M(1 + \|Z_s\|)\,ds = \|z_0\| + 2Mt + \int_0^t 2M\|Z_s\|\,ds\,. \tag{C.1.4}$$

Invoking Grönwall's lemma then yields the result. ∎

The following lemma states that the distance between the "interpolated" and "discretized noise" trajectories is bounded by a factor proportional to the stepsize η.

Lemma C.2. *Fix $\mathcal{K} \subset \mathcal{X}$ compact, $\eta_0 > 0$, $T > 0$. Then, there exists some constant $c = c(\mathcal{K}, \eta_0, T, f, \mathsf{U}, \Omega) < +\infty$ such that, for any $x_0 \in \mathcal{K}$, $\eta \in (0, \eta_0]$, $t \in [0, T]$,*

$$\mathrm{dist}_{[0,T]}(X, Z) \leq c\,\eta\,. \tag{C.1.5}$$

Proof. Before starting, notice that Lemmas B.2 and C.1 imply that there is some compact set $\mathcal{K}'$ that depends on $\mathcal{K}$, T, η_0 and M such that, for any $x_0 \in \mathcal{K}$, $\eta \in (0, \eta_0]$, $t \in [0, T]$, X_t and Z_t belong to $\mathcal{K}'$. In particular, U is therefore Lipschitz-continuous on $\mathcal{K}' \times \Omega$ so that, for any $\omega \in \Omega$, $x \mapsto -\nabla f(x) + \mathsf{U}(x, \omega)$ is Lipschitz-continuous on $\mathcal{K}'$ with constant L and bounded with constant B.

Let us now estimate the derivative of the interpolated trajectory X_t, which is piecewise differentiable by definition (see Eq. (C.1.1b)). For any $t \in [0, T]$ such that $t \in (\eta n, \eta(n+1))$ for some $n \geq 0$, we have that

$$\begin{aligned}
&\left\|\dot{X}_t - \left(-\nabla f(X_t) + \mathsf{U}(X_t, \omega_{\lfloor t/\eta \rfloor})\right)\right\| \\
&\quad = \left\|\frac{x_{n+1} - x_n}{\eta} - \left(-\nabla f(X_t) + \mathsf{U}(X_t, \omega_{\lfloor t/\eta \rfloor})\right)\right\| \\
&\quad = \left\|\left(-\nabla f(X_{\eta n}) + \mathsf{U}(X_{\eta n}, \omega_n)\right) - \left(-\nabla f(X_t) + \mathsf{U}(X_t, \omega_{\lfloor t/\eta \rfloor})\right)\right\| \\
&\quad \leq L\|X_{\eta n} - X_t\|
\end{aligned} \tag{C.1.6}$$

where we used the L-Lipschitz-continuity of $x \mapsto -\nabla f(x) + \mathsf{U}(x, \omega)$ on $\mathcal{K}'$ uniformly in $\omega \in \Omega$. Since this map is also bounded by B, we have that

$$\left\|\dot{X}_t - \left(-\nabla f(X_t) + \mathsf{U}(X_t, \omega_{\lfloor t/\eta \rfloor})\right)\right\| \leq L\|X_{\eta n} - X_t\| \leq L\|x_n - x_{n+1}\| \leq LB\eta\,. \tag{C.1.7}$$

Moreover, the L-Lipschitz-continuity of $x \mapsto -\nabla f(x) + \mathsf{U}(x, \omega)$ also gives us that

$$\begin{aligned}
&\|\dot{Z}_t - \left(-\nabla f(X_t) + \mathsf{U}(X_t, \omega_{\lfloor t/\eta \rfloor})\right)\| \\
&\quad = \|-\nabla f(Z_t) + \mathsf{U}(Z_t, \omega_{\lfloor t/\eta \rfloor}) - \left(-\nabla f(X_t) + \mathsf{U}(X_t, \omega_{\lfloor t/\eta \rfloor})\right)\| \\
&\quad \leq L\|Z_t - X_t\|\,.
\end{aligned} \tag{C.1.8}$$

Putting everything together and integrating $X_t - Z_t$ from 0 to t yields (since X_t is absolutely continuous), for any $t \in [0, T]$,

$$\|X_t - Z_t\| \leq \|X_0 - z_0\| + \int_0^t \|\dot{X}_s - \dot{Z}_s\|\,ds \leq LB\eta + L\int_0^t \|X_s - Z_s\|\,ds \tag{C.1.9}$$

where we used that $z_0 = X_0 = x_0$ by definition of the trajectories. Finally, Grönwall's lemma then yields the result. ∎

C.2. **Large deviation principle for interpolated trajectories.** From the Lagrangian defined in (B.2.11), we define, on $\mathcal{C}_T = \mathcal{C}([0,T],\mathcal{X})$, the normalized action functional $\mathcal{S}_{[0,T]}$ as

$$\mathcal{S}_{[0,T]}(\gamma) = \begin{cases} \int_0^T \mathcal{L}(\gamma_t, \dot\gamma_t)\, dt & \text{if } \gamma \text{ is absolutely continuous,} \\ \infty & \text{otherwise,} \end{cases} \tag{C.2.1}$$

following Freidlin & Wentzell [20, Chap. 3.2], as a manner to quantify how "probable" a trajectory is.

We first show that the set

$$\Gamma^{\mathcal{K}}_{[0,T]}(s) \coloneqq \{\gamma \in \mathcal{C}([0,T],\mathcal{X}) : \gamma_0 \in \mathcal{K}, \mathcal{S}_{[0,T]}(\gamma) \le s\} \tag{C.2.2}$$

of trajectories with bounded action functional is compact and $\mathcal{S}_{[0,T]}$ is lower semi-continuous.

Lemma C.3. *Fix $T > 0$. For any $\mathcal{K} \subset \mathcal{X}$ compact, $s \ge 0$, the set $\Gamma^{\mathcal{K}}_{[0,T]}(s)$ is compact and $\mathcal{S}_{[0,T]}$ is l.s.c. on $\mathcal{C}([0,T],\mathcal{X})$.*

Proof. Let us first check that $\mathcal{S}_{[0,T]}$ is l.s.c. on $\mathcal{C}([0,T],\mathcal{X})$ by applying Ioffe et al. [29, §9.1.4, Thm. 3]: $(t,x,v) \mapsto \mathcal{L}(x,v)$ is a normal integrand since $(x,v) \mapsto \mathcal{L}(x,v)$ is l.s.c. by construction, quasiregular since $\mathcal{L}(x,\cdot)$ is convex for any $x \in \mathcal{X}$ and satisfies the growth condition because it is non-negative (see Lemma B.1). Hence, $\mathcal{S}_{[0,T]}$ is l.s.c. on $\mathcal{C}([0,T],\mathcal{X})$.

The compactness of $\Gamma^{\mathcal{K}}_{[0,T]}(s)$ follows from the idea of proof as Freidlin & Wentzell [20, Chap. 7, Lemma 4.2] but with the added difficulty that the gradient ∇f and the noise U are not uniformly bounded. Take $\gamma \in \Gamma^{\mathcal{K}}_{[0,T]}(s)$. Since $\mathcal{S}_{[0,T]}(\gamma) \le s < +\infty$, it means that $\mathcal{L}(\gamma_t, \dot\gamma_t) < +\infty$ for almost every t so that by Lemma B.1, almost everywhere,

$$\|\dot\gamma_t\| \le 2M(1 + \|\gamma_t\|)\,. \tag{C.2.3}$$

Grönwall's lemma then yields that, for any $t \in [0,T]$,

$$\|\gamma_t\| \le e^{2MT}(\|\gamma_0\| + 2MT) \tag{C.2.4}$$

which is bounded uniformly in $\gamma \in \Gamma^{\mathcal{K}}_{[0,T]}(s)$ since γ_0 is in $\mathcal{K}$ compact. Hence, $\|\dot\gamma_t\|$ is also uniformly bounded by Eq. (C.2.3). Therefore, the functions in $\Gamma^{\mathcal{K}}_{[0,T]}(s)$ are equicontinuous and uniformly bounded and, $\Gamma^{\mathcal{K}}_{[0,T]}(s)$ is closed by l.s.c. of $\mathcal{S}_{[0,T]}$, so that, by the Arzelà-Ascoli theorem, $\Gamma^{\mathcal{K}}_{[0,T]}(s)$ is compact. ∎

The following result establishes the fact that the functional $\eta^{-1}\mathcal{S}_{[0,T]}$ is the action functional in $\mathcal{C}([0,T],\mathcal{X})$ of the interpolated process $(X_t)_{t\in[0,T]}$ of the algorithm started at x_0, uniformly with respect to the initial point x_0 in any compact set $\mathcal{K} \subset \mathcal{X}$, as $\eta \to 0$. This enables to build on Freidlin & Wentzell [20, Chap. 7, Thm. 4.1′] to provide a large deviation principle for the interpolated trajectory $(X_t)_{t\in[0,T]}$, meaning that for η small enough, it will be (a) close to probable trajectories with a probability exponentially big in their action functional; and (b) far from the most probable trajectories with a probability exponentially small in their action value.

Proposition C.1. *Fix $T > 0$. For any $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathcal{X}$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, we have that*

$$\mathbb{P}_{x_0}\big(\mathrm{dist}_{[0,T]}(X,\gamma) < \delta\big) \ge \exp\left(-\frac{\mathcal{S}_{[0,T]}(\gamma) + \varepsilon}{\eta}\right) \tag{C.2.5a}$$

$$\mathbb{P}_{x_0}\left(\mathrm{dist}_{[0,T]}(X, \Gamma^{\{x_0\}}_{[0,T]}(s)) > \delta\right) \le \exp\left(-\frac{s-\varepsilon}{\eta}\right) \tag{C.2.5b}$$

for all $\gamma \in \Gamma^{\{x_0\}}_{[0,T]}(s)$.

Proof. Our strategy is to apply Freidlin & Wentzell [20, Chap. 7, Thm. 4.1′] to the process $(Z_t)_{t\in[0,T]}$ and use Lemma C.2 to transfer the result to $(X_t)_{t\in[0,T]}$ (both starting from x_0). However, this theorem requires $b(x,\omega) \coloneqq -\nabla f(x) + \mathsf{U}(x,\omega)$ to be uniformly bounded as well as its derivative. To avoid this issue, note that, for a fixed compact set $\mathcal{K}$, $s>0$, and all $\eta \le 1$, the trajectories of $\Gamma^{\mathcal{K}}_{[0,T]}(s)$ and of the process $(Z_t)_{t\in[0,T]}$ are contained in a compact set $\mathcal{K}'$ by the compactness of $\Gamma^{\mathcal{K}}_{[0,T]}(s)$ and Lemma C.1.

Now, consider $b'\colon \mathcal{X}\times\mathbb{R}^m \to \mathcal{X}$ twice differentiable that coincides with b on $\mathcal{K}'\times\Omega$ but that is uniformly bounded along with its derivative. Define $\mathcal{H}'(x,v) = \log \mathbb{E}[\exp(\langle v, b'(x,\omega)\rangle)]$ which is still differentiable and satisfies "Condition F" of Freidlin & Wentzell [20, Chap. 7.4] by Freidlin & Wentzell [20, Chap. 7, Lem. 4.3] and i.i.d. assumption. Hence, Freidlin & Wentzell [20, Chap. 7, Thm. 4.1′] yields that, with $\mathcal{K}$ and s fixed above, for any $\delta,\varepsilon>0$, there exists $\eta_0>0$ such that, for any $\eta\in(0,\eta_0]$, for any $x_0\in\mathcal{K}$, $\gamma\in\Gamma^{\{x_0\}}_{[0,T]}(s)$,

$$\mathbb{P}_{x_0}\big(\mathrm{dist}_{[0,T]}(Z,\gamma)<\delta\big) \ge \exp\left(-\frac{\mathcal{S}_{[0,T]}(\gamma)+\varepsilon}{\eta}\right) \tag{C.2.6a}$$

$$\mathbb{P}_{x_0}\left(\mathrm{dist}_{[0,T]}(Z,\Gamma^{\{x_0\}}_{[0,T]}(s))>\delta\right) \le \exp\left(-\frac{s-\varepsilon}{\eta}\right). \tag{C.2.6b}$$

To obtain the result for the process $(X_t)_{t\in[0,T]}$, fix $\delta>0$. By Lemma C.2, there exists $\eta_0>0$ such that, for any $\eta\in(0,\eta_0]$, for any initial point $x_0\in\mathcal{K}$, $\mathrm{dist}_{[0,T]}(X,Z)<\delta$. Combining this with the previous result on $(Z_t)_{t\in[0,T]}$ yields that, for any $\delta,\varepsilon>0$, there exists $\eta_0>0$ such that, for any $\eta\in(0,\eta_0]$, for any $x_0\in\mathcal{K}$, $\gamma\in\Gamma^{\{x_0\}}_{[0,T]}(s)$,

$$\mathbb{P}_{x_0}\big(\mathrm{dist}_{[0,T]}(X,\gamma)<2\delta\big) \ge \exp\left(-\frac{\mathcal{S}_{[0,T]}(\gamma)+\varepsilon}{\eta}\right) \tag{C.2.7a}$$

and

$$\mathbb{P}_{x_0}\left(\mathrm{dist}_{[0,T]}(X,\Gamma^{\{x_0\}}_{[0,T]}(s))>2\delta\right) \le \exp\left(-\frac{s-\varepsilon}{\eta}\right) \tag{C.2.7b}$$

which concludes the proof. ∎

C.3. **Large deviation principle for discrete trajectories.** We now leverage the previous section to show a LDP for the discrete rescaled trajectories $(x^\eta_n)_{n\ge0}$ defined in Eq. (C.1.1a). To do so, we will use the results of the previous section by considering a finite number of points $(X_n)_{n\ge0}$ from the continuous interpolation.

For some $N>0$, we will first equip $\mathcal{X}^N$ with the distance

$$\mathrm{dist}_N(\xi,\zeta) = \max_{0\le n\le N-1} \|\xi_n-\zeta_n\| \tag{C.3.1}$$

and bound the difference between the discrete rescaled trajectory and the continuous interpolation.

Lemma C.4. *Fix $\mathcal{K}\subset\mathcal{X}$ compact, $\eta_0>0$, $N\ge1$. Then, there exists some constant $c=c(\mathcal{K},\eta_0,N,f,\mathsf{U},\Omega)<+\infty$ such that, for any $x_0\in\mathcal{K}$, $\eta\in(0,\eta_0]$,*

$$\mathrm{dist}_N(x^\eta,(X_n)_{0\le n\le N-1}) = \max_{0\le n\le N-1}\|x^\eta_n - X_n\| \le c\eta\,. \tag{C.3.2}$$

Proof. By Lemma B.2, there is some compact set $\mathcal{K}'$ that depends on $\mathcal{K}$, N, η_0 and M such that, for any $x_0\in\mathcal{K}$, $\eta\in(0,\eta_0]$, $n\le\lceil 1/\eta\rceil N$, x_n belongs to $\mathcal{K}'$. In particular, for almost every $t\in[0,N-1]$, we have, that,

$$\dot X_t = \frac{x_{\lfloor t/\eta\rfloor+1}-x_{\lfloor t/\eta\rfloor}}{\eta} = -\nabla f(x_{\lfloor t/\eta\rfloor}) + \mathsf{U}(x_{\lfloor t/\eta\rfloor},\omega_{\lfloor t/\eta\rfloor}) \tag{C.3.3}$$

with $x_{\lfloor t/\eta \rfloor}$ belonging to $\mathcal{K}'$ since $\left\lfloor \frac{t}{\eta} \right\rfloor \leq \left\lfloor \frac{N-1}{\eta} \right\rfloor \leq \left\lceil \frac{N}{\eta} \right\rceil$. Hence, the norm of $\dot{X}_t$ is bounded by some constant B for almost every $t \in [0, N-1]$ uniformly in $x_0 \in \mathcal{K}$, $\eta \in (0, \eta_0]$ by the growth condition in (*b*).

Therefore, for any $0 \leq n \leq N-1$, since $x_n^\eta = x_{n\lfloor 1/\eta \rfloor} = X_{n\eta\lfloor 1/\eta \rfloor}$, we have that

$$\|x_n^\eta - X_n\| \leq Bn|1 - \eta\lfloor 1/\eta \rfloor| \leq BN\eta \tag{C.3.4}$$

which concludes the proof. ∎

Now, for $N \geq 0$, $\xi = (\xi_0, \dots, \xi_{N-1}) \in \mathcal{X}^N$, let us define the normalized discrete action functional

$$\mathcal{A}_N(\xi) \coloneqq \sum_{n=0}^{N-2} \rho(\xi_n, \xi_{n+1}) \tag{C.3.5}$$

where the cost of moving from one iteration to the next is defined for any $x, x' \in \mathcal{X}$ from the previous continuous normalized action functional (cf. Eq. (C.2.1)) with horizon 1 as

$$\rho(x, x') \coloneqq \inf\{\mathcal{S}_{0,1}(\gamma) : \gamma \in \mathcal{C}([0,1], \mathcal{X}), \gamma_0 = x, \gamma_1 = x'\}. \tag{C.3.6}$$

Again, we show that the set of discrete trajectories with low action functional

$$\Gamma_N^{\mathcal{K}}(s) \coloneqq \{\xi \in \mathcal{X}^N : \xi_0 \in \mathcal{K}, \mathcal{A}_N(\xi) \leq s\} \tag{C.3.7}$$

is compact and $\mathcal{A}_N$ is l.s.c..

Lemma C.5. *Fix $N \geq 0$. For any $\mathcal{K} \subset \mathcal{X}$ compact, $s \geq 0$, the set $\Gamma_N^{\mathcal{K}}(s)$ is compact and $\mathcal{A}_N$ is l.s.c. on $\mathcal{X}^N$.*

Proof. First, let us show that, for $s \geq 0$, $\mathcal{K} \subset \mathcal{X}$, the set $\{(x, x') \in \mathcal{K} \times \mathcal{X} : \rho(x, x') \leq s\}$ is compact. Let $(x^k, x'^k)_{k \geq 0}$ be a sequence in $\mathcal{K} \times \mathcal{X}$ such that $\rho(x^k, x'^k) \leq s$ for all $k \geq 0$. Then, for any $k \geq 0$, there exists $\gamma^k \in \mathcal{C}([0,1], \mathcal{X})$ such that $\gamma_0^k = x^k$, $\gamma_1^k = x'^k$ and $\mathcal{S}_{0,1}(\gamma^k) \leq \rho(x^k, x'^k) + (1+k)^{-1}$. In particular, for any $k \geq 0$, γ^k belongs to $\{\gamma \in \mathcal{C}([0,1], \mathcal{X}) : \gamma_0 \in \mathcal{K}, \mathcal{S}_{0,1}(\gamma) \leq s+1\}$ which is compact by Lemma C.3. Hence, there exists a subsequence that converges uniformly to some $\gamma \in \mathcal{C}([0,1], \mathcal{X})$. Without loss of generality, assume that $\gamma^k \to \gamma$ as $k \to \infty$. In particular, (x^k, x'^k) converges to (γ_0, γ_1) as $k \to \infty$ with γ_0 belonging to $\mathcal{K}$. Then, by l.s.c. of $\mathcal{S}_{0,1}$, we have that $\mathcal{S}_{0,1}(\gamma) \leq \liminf_{k \to \infty} \rho(x^k, x'^k) \leq s$ so that $\rho(\gamma_0, \gamma_1) \leq s$. Therefore, we have shown that $\{(x, x') \in \mathcal{K} \times \mathcal{X} : \rho(x, x') \leq s\}$ is compact.

As a consequence ρ is l.s.c. on $\mathcal{X} \times \mathcal{X}$: indeed, for any $s \geq 0$, any convergence sequence of points of $\{(x, x') \in \mathcal{X} \times \mathcal{X} : \rho(x, x') \leq s\}$ must be included in $\{(x, x') \in \mathcal{K} \times \mathcal{X} : \rho(x, x') \leq s\}$ for some $\mathcal{K} \subset \mathcal{X}$ compact, which is closed and therefore contains the limit point of any such sequences. $\mathcal{A}_N$ is then immediately l.s.c. on $\mathcal{X}^N$.

Finally, the compactness of $\Gamma_N^{\mathcal{K}}(s)$ follows by induction on N. For $N = 1$, $\Gamma_1^{\mathcal{K}}(s) = \{(x, x') \in \mathcal{X} \times \mathcal{X} : \rho(x, x') \leq s\}$ which is compact by the previous argument for any compact set $\mathcal{K}$. Now, assume that $\Gamma_N^{\mathcal{K}}(s)$ is compact for some $N \geq 1$. As a consequence, the set

$$\mathcal{K}' \coloneqq \{x \in \mathcal{X} : \exists (\xi_0, \dots, \xi_{N-2}) \in \mathcal{X}^{N-1}, (\xi_0, \dots, \xi_{N-2}, x) \in \Gamma_N^{\mathcal{K}}(s)\} \tag{C.3.8}$$

is compact as well. Hence, $\Gamma_{N+1}^{\mathcal{K}}(s)$ is included in teh product of compact sets $\Gamma_{N-1}^{\mathcal{K}}(s) \times \Gamma_1^{\mathcal{K}'}(s)$ and is therefore bounded. Moreover, $\Gamma_{N+1}^{\mathcal{K}}(s)$ is closed by l.s.c. of $\mathcal{A}_N$ and therefore compact. This concludes the proof by induction and the proof of the lemma. ∎

We will first provide a discrete analogue of Proposition C.1 for the interpolated trajectory at times $n = 0, \dots, N-1$, with $\eta^{-1}\mathcal{A}_N$ the action functional in $\mathcal{X}^N$ of the process $(X_n)_{0 \leq n \leq N-1}$, uniformly with respect to the starting point x_0 in any compact set $\mathcal{K} \subset \mathcal{X}$, as $\eta \to 0$.

Proposition C.2. *Fix $N \geq 0$. For any $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathcal{X}$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $\xi \in \Gamma_N^{\{x_0\}}(s)$, we have that*

$$\mathbb{P}_{x_0}(\mathrm{dist}_N((X_n)_{0\leq n\leq N-1}, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{A}_N(\xi) + \varepsilon}{\eta}\right) \tag{C.3.9a}$$

and

$$\mathbb{P}_{x_0}\left(\mathrm{dist}_N((X_n)_{0\leq n\leq N-1}, \Gamma_N^{\{x_0\}}(s)) > \delta\right) \leq \exp\left(-\frac{s-\varepsilon}{\eta}\right). \tag{C.3.9b}$$

for all $\xi \in \Gamma_N^{\{x_0\}}(s)$

Proof. Invoke the first part of Proposition C.1 with $T \leftarrow N-1$ and $s \leftarrow s + \varepsilon$. There exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $\gamma \in \Gamma_{0,N}^{\{x_0\}}(s+\varepsilon)$,

$$\mathbb{P}_{x_0}(\mathrm{dist}_{0,N-1}(X, \gamma) < \delta) \geq \exp\left(-\frac{\mathcal{S}_{0,N-1}(\gamma) + \varepsilon}{\eta}\right). \tag{C.3.10}$$

Take $\eta \in (0, \eta_0]$, $x_0 \in \mathcal{K}$, $\xi \in \Gamma_N^{\{x_0\}}(s)$. Then, there exists $\gamma \in \Gamma_{0,N}^{x_0}(s+\varepsilon)$, for any $0 \leq n \leq N-1$, $\gamma_n = \xi_n$, cf. (C.3.6). Hence,

$$\begin{aligned}
\mathbb{P}_{x_0}(\mathrm{dist}_N((X_n)_{0\leq n\leq N-1}, \xi) < \delta) &\geq \mathbb{P}_{x_0}(\mathrm{dist}_{0,N-1}(X, \gamma) < \delta) \\
&\geq \exp\left(-\frac{\mathcal{S}_{0,N-1}(\gamma) + \varepsilon}{\eta}\right) \\
&\geq \exp\left(-\frac{\mathcal{A}_N(\xi) + 2\varepsilon}{\eta}\right)
\end{aligned} \tag{C.3.11}$$

which prove the first part of the result.

For the second part, we have similarly from Proposition C.1 with $T \leftarrow N-1$ and $\delta \leftarrow \delta/2$ that for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$,

$$\mathbb{P}_{x_0}\left(\mathrm{dist}_{0,N-1}(X, \Gamma_{0,N-1}^{\{x_0\}}(s)) > \delta/2\right) \leq \exp\left(-\frac{s-\varepsilon}{\eta}\right). \tag{C.3.12}$$

Now, note that if $\mathrm{dist}_{0,N-1}(X, \Gamma_{0,N-1}^{\{x_0\}}(s)) < \delta$, then there must exist $\gamma \in \Gamma_{0,N-1}^{\{x_0\}}(s)$ such that $\mathrm{dist}_{0,N-1}(X, \gamma) \leq \delta$. Consider the discrete path $\xi \in \mathcal{X}^N$ defined by $\xi_n = \gamma_n$ for $0 \leq n \leq N-1$. Then, by construction, $\mathcal{A}_N(\xi) \leq \mathcal{S}_{0,N-1}(\gamma) \leq s$ and $\mathrm{dist}_N((X_n)_{0\leq n\leq N-1}, \xi) \leq \delta$. Thus,

$$\mathrm{dist}_{0,N-1}(X, \Gamma_{0,N-1}^{\{x_0\}}(s)) < \delta \implies \mathrm{dist}_N((X_n)_{0\leq n\leq N-1}, \Gamma_N^{\{x_0\}}(s)) \leq \delta\,. \tag{C.3.13}$$

Putting all together, we have that

$$\begin{aligned}
\mathbb{P}_{x_0}\left(\mathrm{dist}_N((X_n)_{0\leq n\leq N-1}, \Gamma_N^{\{x_0\}}(s)) > \delta\right) &\leq \mathbb{P}_{x_0}\left(\mathrm{dist}_{0,N-1}(X, \Gamma_{0,N-1}^{\{x_0\}}(s) \geq \delta)\right) \\
&\leq \mathbb{P}_{x_0}\left(\mathrm{dist}_{0,N-1}(X, \Gamma_{0,N-1}^{\{x_0\}}(s) > \delta/2)\right) \\
&\leq \exp\left(-\frac{s-\varepsilon}{\eta}\right)
\end{aligned} \tag{C.3.14}$$

which concludes our proof. ∎

Finally, we end up with a large deviation principle on the discrete rescaled iterates $(x_n^\eta)_{0\leq n\leq N-1} = (x_{n\lfloor \eta^{-1} \rfloor})_{0\leq n}$ by leveraging Lemma C.4. In the following result, the functional $\eta^{-1}\mathcal{A}_N$ is thus the action functional in $\mathcal{X}^N$ of the process $(x_n^\eta)_{0\leq n\leq N-1}$ uniformly with respect to the starting point x_0 in any compact set $\mathcal{K} \subset \mathcal{X}$, as $\eta \to 0$.

Corollary C.1. *Fix $N \geq 0$. For any $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathcal{X}$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $\xi \in \Gamma_N^{\{x_0\}}(s)$, we have that*

$$\mathbb{P}_{x_0}(\mathrm{dist}_N(x^\eta, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{A}_N(\xi) + \varepsilon}{\eta}\right) \tag{C.3.15a}$$

and

$$\mathbb{P}_{x_0}\left(\mathrm{dist}_N(x^\eta, \Gamma_N^{\{x_0\}}(s)) > \delta\right) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right). \tag{C.3.15b}$$

for all $\xi \in \Gamma_N^{\{x_0\}}(s)$.

Proof. Fix $\delta > 0$. Choose η_0 such that both Proposition C.2 and Lemma C.4 hold with $c\eta \leq \delta$. Then for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $\xi \in \Gamma_N^{\{x_0\}}(s)$, $\xi \in \Gamma_N^{\{x_0\}}(s)$, we have that

$$\mathbb{P}_{x_0}(\mathrm{dist}_N((X_n)_{0 \leq n \leq N-1}, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{A}_N(\xi) + \varepsilon}{\eta}\right). \tag{C.3.16}$$

Now, if $\mathrm{dist}_N((X_n)_{0 \leq n \leq N-1}, \xi) < \delta$ and $\mathrm{dist}_N(x^\eta, (X_n)_{0 \leq n \leq N-1}) \leq \delta$, then $\mathrm{dist}_N(x^\eta, \xi) < 2\delta$. Thus,

$$\mathbb{P}_{x_0}(\mathrm{dist}_N(x^\eta, \xi) < 2\delta) \geq \mathbb{P}_{x_0}(\mathrm{dist}_N((X_n)_{0 \leq n \leq N-1}, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{A}_N(\xi) + \varepsilon}{\eta}\right). \tag{C.3.17}$$

The second part can be obtained similarly. ∎

To summarize this part on large deviations principles, we state a corollary containing all the results that will be needed in the following sections.

Corollary C.2. *Fix $N \geq 0$. Then:*

- *For all $s > 0$, the set*

$$\Gamma_N^{\mathcal{K}}(s) \coloneqq \{\xi \in \mathcal{X}^N : \xi_0 \in \mathcal{K}, \mathcal{A}_N(\xi) \leq s\} \tag{C.3.18}$$

  is compact and $\mathcal{A}_N$ is l.s.c. on $\mathcal{X}^N$.

- *For all $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathcal{X}$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for all $x_0 \in \mathcal{K}$, $n \leq N$, $\xi \in \Gamma_n^{\{x_0\}}(s)$, we have that*

$$\mathbb{P}_{x_0}(\mathrm{dist}_n(x^\eta, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{A}_n(\xi) + \varepsilon}{\eta}\right) \tag{C.3.19a}$$

  and

$$\mathbb{P}_{x_0}\left(\mathrm{dist}_n(x^\eta, \Gamma_n^{\{x_0\}}(s)) > \delta\right) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right). \tag{C.3.19b}$$

Appendix D. Attractors and limiting measures via large deviations

We now take inspiration from the framework of Kifer [35] in order to relate the sets of critical points to the sets where points can move at no cost. Then, we relate the probability of SGD moving to neighborhoods of critical sets to the probability of being close to well-chosen paths, which enables us to use the results of the previous section. Finally, we build upon these results to provide bounds on the limiting measure of SGD.

D.1. **Setup.** We first need to define the gradient flow of f.

Definition 2. Define, for $x \in \mathcal{X}$, the flow Θ of $-\nabla f$ starting at x, i.e.,

$$\dot{\Theta}_t(x) = -\nabla f(\Theta_t(x)) \quad \text{with} \quad \Theta_0(x) = x \tag{D.1.1}$$

and let $F(x)$ be the value of this flow at time 1, i.e.,

$$F(x) = \Theta_1(x)\,. \tag{D.1.2}$$

Lemma D.1 (Properties of the flow)**.** *Θ is well-defined and continous in both time and space, and, for any $T \geq 0$, $\gamma \in \mathcal{C}([0,T],\mathcal{X})$ such that $\gamma_0 = x$,*

$$\mathcal{S}_{0,T}(\gamma) = 0 \iff \gamma_t = \Theta_t(x) \;\; \textit{for all } t \in [0,T]\,. \tag{D.1.3}$$

Proof. The well-definition and continuity of Θ are a consequence of f being twice continuously differentiable and of the global Cauchy-Lipschitz (Picard–Lindelöf) theorem for ordinary differential equation (ODE). The equivalence follows from the uniqueness of the flow and Lemma B.1 since

$$\begin{aligned} \mathcal{S}_{0,T}(\gamma) = 0 &\iff \mathcal{L}(\gamma_t, \dot{\gamma}_t) = 0 \text{ almost everywhere} \\ &\iff \dot{\gamma}_t = -\nabla f(\gamma_t) \text{ almost everywhere} \end{aligned} \tag{D.1.4}$$

and thus, by extending $\dot{\gamma}$ by continuity, both γ and Θ satisfy the same ODE with the same initial condition and are thus equal for all t by uniqueness of the solutions. ∎

The following lemma translates this for F.

Lemma D.2 (Properties of F)**.** *F is well-defined and continuous and, for any $x, x' \in \mathcal{X}$,*

$$\rho(x,x') = 0 \iff x' = F(x)\,. \tag{D.1.5}$$

Proof. The implication ($\Leftarrow$) is immediate by definition of F and Lemma D.1. Now for the reverse, assume that $\rho(x,x') = 0$. Following the proof of Lemma D.1, there exists $\gamma \in \mathcal{C}([0,1],\mathcal{X})$ such that $\gamma_0 = x$, $\gamma_1 = x'$ and $\mathcal{S}_{0,1}(\gamma) = 0$. By Lemma D.1, $\gamma_t = \Theta_t(x)$ for all $t \in [0,1]$ and thus $x' = \Theta_1(x) = F(x)$. ∎

D.2. **Attractors.** Let us first formalize the minimum-energy displacement between two points.

Definition 3 (Kifer [35, §1.5])**.** Define, for $x, x' \in \mathcal{X}$,

$$\begin{aligned} B(x,x') &= \inf\left\{\mathcal{S}_{[0,T]}(\gamma) : \gamma \in \mathcal{C}([0,T],\mathcal{X}), \gamma_0 = x, \gamma_T = x', T \in \mathbb{N}, T \geq 1\right\} \\ &= \inf\left\{\mathcal{A}_N(\xi) : \xi \in \mathcal{X}^N\,,\, \xi_0 = x\,,\, \xi_{N-1} = x',\, N \geq 1\right\}. \end{aligned} \tag{D.2.1}$$

The fact that these two expressions coincide directly come from the definition of ρ.

This enables us to define an equivalence relation for the critical points of f by grouping points connected by a null-energy path.

Proposition D.1. *The relation $\sim$ defined for any $x, x' \in \operatorname{crit} f$ as*

$$x \sim x' \iff B(x,x') = B(x',x) = 0 \tag{D.2.2}$$

is an equivalence relation on $\operatorname{crit} f$.

Proof.
- Reflexivity: $B(x,x) = 0$ by Lemma D.1 since the flow started at $x \in \operatorname{crit} f$ is constant.
- Symmetry: this follows from the definition of $\sim$.
- Transitivity: for any $x, x', x'' \in \operatorname{crit} f$, we have by construction of B that

$$B(x,x'') \leq B(x,x') + B(x',x'')\,. \tag{D.2.3}$$

Therefore, if $x \sim x'$ and $x' \sim x''$, then $B(x,x'') = 0$. $B(x'',x) = 0$ follows with a symmetric argument and thus $x \sim x''$. ∎

Near critical points of f, the Lagrangian $\bar{\mathcal{L}}$ is actually very regular.

Lemma D.3. *For any $x \in \mathcal{X}$, there exists $\delta > 0$ such that $\bar{\mathcal{L}}$ is finite and jointly Lipschitz continuous on $\mathbb{B}(x,\delta) \times \mathbb{B}(0,\delta)$.*

Moreover, the following supremum is finite:

$$\sup\left\{ \frac{\mathcal{L}(x',v)}{\|v\|^2} : x' \in \mathbb{B}(x,\delta) \cap \operatorname{crit} f\,, v \in \mathbb{B}(0,\delta) \right\} < \infty\,. \tag{D.2.4}$$

Proof. Take $x \in \mathcal{X}$. We apply the implicit function theorem to the equation

$$\nabla_v \bar{\mathcal{H}}(x',w) = v\,, \tag{D.2.5}$$

in the variables $(x', v, w) \in \mathcal{X} \times \mathbb{R}^d \times \mathbb{R}^d$.

We derive that

$$\nabla_v \bar{\mathcal{H}}(x',v) = \frac{\mathbb{E}[\mathsf{U}(x',\omega)\exp(\langle v, \mathsf{U}(x',\omega)\rangle)]}{\mathbb{E}[\exp(\langle v, \mathsf{U}(x',\omega)\rangle)]} \tag{D.2.6}$$

and thus $(x,0,0)$ is solution of (D.2.5) since

$$\nabla_v \bar{\mathcal{H}}(x,0) = \mathbb{E}[\mathsf{U}(x,\omega)] = 0\,. \tag{D.2.7}$$

Moreover, $\operatorname{Hess}_v \bar{\mathcal{H}}(x,0) = \mathbb{E}[\mathsf{U}(x,\omega)\mathsf{U}(x,\omega)^\top]$ which is positive definite and thus invertible by the blanket assumptions.

Hence, we can apply the implicit function theorem to get that there exists $\delta > 0$, $w\colon \mathbb{B}(x,\delta) \times \mathbb{B}(0,\delta) \to \mathbb{R}^d$ $\mathcal{C}^2$ such that, for any $x' \in \mathbb{B}(x,\delta)$, $v \in \mathbb{B}(0,\delta)$,

$$\nabla_v \bar{\mathcal{H}}(x', w(x',v)) = v\,. \tag{D.2.8}$$

Therefore, for any $x' \in \mathbb{B}(x,\delta)$, $v \in \mathbb{B}(0,\delta)$, since $\bar{\mathcal{L}}(x',v) = \bar{\mathcal{H}}(x',\cdot)^*(v)$, we have

$$\bar{\mathcal{L}}(x',v) = \langle v, w(x',v)\rangle - \bar{\mathcal{H}}(x', w(x',v))\,, \tag{D.2.9}$$

which is finite and $\mathcal{C}^2$ on $\mathbb{B}(x,\delta) \times \mathbb{B}(0,\delta)$. Therefore, $\bar{\mathcal{L}}$ is actually L-jointly Lipschitz continuous on $\overline{\mathbb{B}}(x,\delta/2) \times \overline{\mathbb{B}}(0,\delta/2)$.

For the second part of the lemma, note that the implicit function theorem also ensures that there is $\mathcal{V} \subset \mathbb{R}^d$ a neighborhood of 0 such that, for any $x' \in \mathbb{B}(x,\delta)$, $v \in \mathbb{B}(0,\delta)$, $w(x',v)$ is the unique solution of Eq. (D.2.5) in $\mathcal{V}$. But, for any $x' \in \mathbb{B}(x,\delta)$ and $v = 0$, $w = 0$ is a solution of Eq. (D.2.5) in $\mathcal{V}$ so that necessarily $w(x',0) = 0$, and, as a consequence, $\nabla_v \bar{\mathcal{L}}(x',0) = 0$.

Hence, for any $x' \in \overline{\mathbb{B}}(x,\delta/2)$, $v \in \overline{\mathbb{B}}(0,\delta/2)$,

$$\begin{aligned} \bar{\mathcal{L}}(x',v) &= \bar{\mathcal{L}}(x',v) - \bar{\mathcal{L}}(x',0) - \langle v, \nabla_v \bar{\mathcal{L}}(x',0)\rangle \\ &\le \frac{1}{2} \sup_{\overline{\mathbb{B}}(x,\delta/2)\times\overline{\mathbb{B}}(0,\delta/2)} \left\|\operatorname{Hess}_v \bar{\mathcal{L}}\right\| \|v\|^2\,. \end{aligned} \tag{D.2.10}$$

To conclude, it suffices to note that, for any $x' \in \mathbb{B}(x,\delta) \cap \operatorname{crit} f$, $\nabla f(x) = 0$ and therefore $\mathcal{L}(x',\cdot)$ and $\bar{\mathcal{L}}(x',\cdot)$ coincide. ∎

Lemma D.4. *There exists an open neighborhood $\mathcal{N} \subset (\mathcal{X})^2$ of $(\operatorname{crit} f)^2$ such that ρ is finite and continuous on $\mathcal{N}$.*

Proof. Take $x \in \mathcal{X}$ such that $\nabla f(x) = 0$. We show that there exists a neighborhood of (x,x) on which ρ is finite and continuous.

By Lemma D.3, there exists $\delta > 0$ such that $\bar{\mathcal{L}}$ is finite and L-jointly Lipschitz continuous on $\mathbb{B}(x,\delta) \times \mathbb{B}(0,\delta)$. In particular, for any $x' \in \mathbb{B}(x,\delta)$, $v \in \mathbb{B}(0,\delta)$,

$$\bar{\mathcal{L}}(x',v) \le L\|v\|\,. \tag{D.2.11}$$

By continuity of ∇f, there is $\delta' > 0$, $\delta' < \delta$ such that, for every $x' \in \mathbb{B}(x, \delta')$, $\|\nabla f(x')\| \leq \delta/4$. Then, for any $x' \in \mathbb{B}(x, \delta')$, $v \in \mathbb{B}(0, \delta/2)$,

$$\begin{aligned}\mathcal{L}(x', v) &= \bar{\mathcal{L}}(x', v + \nabla f(x')) \\ &\leq L(\|v + \nabla f(x')\|) \\ &\leq L(\|v\| + \delta/4)\,. \end{aligned} \tag{D.2.12}$$

Take x_i' in $\mathbb{B}(x, \delta')$ for $i = 1, \ldots, 4$ and $\varepsilon > 0$. By definition of ρ, there exists $\gamma \in \mathcal{C}([0,1], \mathcal{X})$ such that $\gamma_0 = x_1'$, $\gamma_1 = x_2'$, $\mathcal{S}_{0,1}(\gamma) \leq \rho(x_1', x_2') + \varepsilon$.

For $0 < s_1 < 1$ small enough and $0 < s_2 < 1$ close enough to 1, we have that γ_t belongs to $\mathbb{B}(x, \delta')$ for any $t \in [0, s_1] \cup [s_2, 1]$. Now we can define $\varphi \in \mathcal{C}([0,1], \mathcal{X})$ that connects x_3' to x_4' by

$$\varphi_t = \begin{cases} x_3' + (t/s_1)(\gamma_{s_1} - x_3') & \text{if } t \in [0, s_1], \\ \gamma_t & \text{if } t \in [s_1, s_2] \\ \gamma_{s_2} + ((t - s_2)/(1 - s_2))(x_4' - \gamma_{s_2}) & \text{if } t \in [s_2, 1]. \end{cases} \tag{D.2.13}$$

Since φ is a continuous path between x_3' and x_4', we have that

$$\rho(x_3', x_4') \leq \mathcal{S}_{0,1}(\varphi)\,. \tag{D.2.14}$$

For s_1 small enough and s_2 close enough to 1, φ belongs to $\mathbb{B}(x, \delta')$ on $[0, s_1] \cup [s_2, 1]$ and thus, its cost can be bounded as

$$\mathcal{S}_{0,1}(\varphi) \leq L(\|\gamma_{s_1} - x_3'\| + s_1\delta/4) + \mathcal{S}_{0,1}(\gamma) + L(\|\gamma_{s_2} - x_4'\| + (1 - s_2)\delta/4) \tag{D.2.15}$$

where we used that γ_{s_1}, γ_{s_2}, x_3', x_4' all belong to $\mathbb{B}(x, \delta')$.

Now, take (x_3', x_4') sufficient close to (x_1', x_2'), and s_1 small enough and s_2 close enough to 1, so that $L(\|\gamma_{s_1} - x_3'\| + s_1\delta/4) + L(\|\gamma_{s_2} - x_4'\| + (1 - s_2)\delta/4) \leq \varepsilon$. Putting everything together yields that

$$\rho(x_3', x_4') \leq \mathcal{S}_{0,1}(\varphi) \leq \mathcal{S}_{0,1}(\gamma) + \varepsilon \leq \rho(x_1', x_2') + 2\varepsilon\,. \tag{D.2.16}$$

Exchanging the roles of (x_1', x_2') and (x_3', x_4') yields the reverse inequality and thus, ρ is continuous on $\mathbb{B}(x, \delta)^2$. ∎

The following lemma relates the Lagrangian to the gradient and our noise structure.

Lemma D.5. *For any $x \in \mathcal{X}$, $v \in \mathbb{R}^d$,*

$$\mathcal{L}(x, v) \geq \frac{\|v + \nabla f(x)\|^2}{2\sigma_\infty^2(f(x))}\,. \tag{D.2.17}$$

Proof. For any $x \in \mathcal{X}$, $v \in \mathbb{R}^d$, we have that

$$\mathcal{H}(x, v) = -\langle v, \nabla f(x)\rangle + \bar{\mathcal{H}}(x, v) \leq -\langle v, \nabla f(x)\rangle + \frac{1}{2}\sigma_\infty^2(f(x))\|v\|^2\,. \tag{D.2.18}$$

Taking the conjugate then yields that

$$\mathcal{L}(x, v) \geq \frac{\|v + \nabla f(x)\|^2}{2\sigma_\infty^2(f(x))}\,. \tag{D.2.19}$$

∎

Now, let us define a potential function U_∞ on $\mathcal{X}$ that uses the minimal displacement energy between two points that will be heavily used in the proofs.

Definition 4 (Potential)**.** Define, for $x \in \mathcal{X}$

$$U_\infty(x) = 2\alpha_\infty(f(x)) \tag{D.2.20}$$

where $\alpha_\infty : \mathcal{X} \to \mathbb{R}$ is a twice continuously differentiable primitive of $1/\sigma_\infty^2$.

Lemma D.6. *For any $x, x' \in \mathcal{X}$,*

$$U_\infty(x') - U_\infty(x) \leq 2B(x, x')\,. \tag{D.2.21}$$

Proof. By Definition 3, there exists $T \geq 1$, $\gamma \in \mathcal{C}([0,T], \mathcal{X})$ such that $\gamma_0 = x$, $\gamma_T = x'$ and $\mathcal{S}_{[0,T]}(\gamma) \leq B(x, x') + \varepsilon$. Then, we have that,

$$\begin{aligned} U_\infty(x') - U_\infty(x) &= 2\int_0^T \frac{\langle \dot\gamma_t, \nabla f(\gamma_t)\rangle}{\sigma_\infty^2(f(\gamma_t))}\,dt \\ &\leq \int_0^T \frac{\|\dot\gamma_t + \nabla f(\gamma_t)\|^2}{\sigma_\infty^2(f(\gamma_t))}\,dt \\ &\leq \int_0^T \mathcal{L}(\gamma_t, \dot\gamma_t)\,dt = 2\mathcal{S}_{[0,T]}(\gamma) \end{aligned} \tag{D.2.22}$$

where we used Lemma D.5 in the last inequality. Finally, out choice of γ implies that

$$U_\infty(x') - U_\infty(x) \leq 2(B(x, x') + \varepsilon) \tag{D.2.23}$$

which concludes the proof. ∎

Lemma D.7 (Equivalence classes are closed)**.** *Equivalence classes of $\sim$ are closed in $\mathcal{X}$. As a consequence, equivalence classes are compact.*

Proof. Let $x \in \operatorname{crit} f$, and take any sequence $(x'_k)_{k\geq 0}$ in $\operatorname{crit} f$ such that $x \sim x'_k$ for every $k \geq 0$ and which converges to some $x' \in \mathcal{X}$. To show that the equivalence classes are closed, we need to show that $x \sim x'$; then compacity follow directly since the equivalence classes are subsets of $\operatorname{crit} f$ which is compact by assumption.

Since $\operatorname{crit} f$ is closed, it holds that x' belongs to $\operatorname{crit} f$. We now show that both $B(x, x')$ and $B(x', x)$ are null. We only show that $B(x, x') = 0$ since the proof for the other equality is symmetric.

As x' is a critical point of f, we have by Lemma D.4, $\rho(\cdot, x')$ is finite and continuous on a neighborhood of x'. Moreover, since x' is a critical point, we have $x' = F(x')$ (see Eq. (D.1.2)) and thus $\rho(x', x') = \rho(x', F(x')) = 0$ by Lemma D.2. Therefore, for any $\varepsilon > 0$ there is a neighborhood of x' on which $\rho(\cdot, x') \leq \varepsilon$. Take k large enough so that x'_k belongs to this neighborhood. Since $x \sim x'_k$, there exists $N \geq 1$, $\xi \in (\mathcal{X})^N$ such that $\xi_0 = x$, $\xi_{N-1} = x'_k$ and $\mathcal{A}_N(\xi) \leq \varepsilon$. Then, the path $\zeta := (\xi_0, \dots, \xi_{N-1}, x') \in (\mathcal{X})^{N+1}$ and satisfies $\zeta_0 = x$, $\zeta_N = x'$ and

$$\mathcal{A}_{N+1}(\zeta) \leq \mathcal{A}_N(\xi) + \rho(x'_k, x') \leq 2\varepsilon\,. \tag{D.2.24}$$

Hence, we have shown that, for any $\varepsilon > 0$, $B(x, x') \leq 2\varepsilon$ so that $B(x, x') = 0$. ∎

Lemma D.8. *For any set $\mathcal{C} \subset \operatorname{crit} f$, there is $r_0 > 0$ such that, for any $0 < r \leq r_0$,*

$$\mathcal{W}_r(\mathcal{C}) := \{x \in \mathcal{X} : \rho(x, \mathcal{C}) < r,\ \rho(\mathcal{C}, x) < r\} \tag{D.2.25}$$

is open and contains $\mathcal{C}$.

Proof. For any $x \in \mathcal{C}$, $\nabla f(x) = 0$ and therefore x is a fixed point of Θ. Lemma D.2 then implies that $\rho(x, x) = 0$. Hence, $\mathcal{W}_r(\mathcal{C})$ indeed contains $\mathcal{C}$. The fact that $\mathcal{W}_r(\mathcal{C})$ is open for $r > 0$ small enough follows from the continuity of ρ close to $\operatorname{crit} f \times \operatorname{crit} f$ (Lemma D.4), which is compact. ∎

This lemma is adapted and significantly expanded from Kifer [35, §1.5, Lem. 5.2] to handle both the unboundedness of the space and the fact that B is neither l.s.c. nor upper semi-continuous (u.s.c.).

Lemma D.9. *Let $\mathcal{K}$ be an equivalence class of $\sim$. Then, for any $\varepsilon > 0$, there is some $N \geq 1$ such that, for any $x, z \in \mathcal{K}$, there is $\xi \in (\mathcal{X})^N$ such that $\xi_0 = x$, $\xi_{N-1} = z$, $\mathcal{A}_N(\xi) < \varepsilon$ and $\max_{0 \leq n < N} d(\xi_n, \mathcal{K}) < \varepsilon$.*

Proof. By Lemma D.7, $\mathcal{K}$ is a compact set. Moreover, $\mathcal{K}$ is made of critical points of f so that by Lemma D.4, ρ is finite and continuous on a neighborhood of $\mathcal{K} \times \mathcal{K}$. By compactness of $\mathcal{K} \times \mathcal{K}$, ρ is actually uniformly continuous on a neighborhood of $\mathcal{K} \times \mathcal{K}$ so, in particular, for any $\varepsilon > 0$, there exists $\delta > 0$ such that, for any $x_i \in \mathcal{X}$ with $d(x_i, \mathcal{K}) < \delta$ for $i = 1, \dots, 4$ such that $\|x_1 - x_2\| < \delta$, $\|x_3 - x_4\| < \delta$,

$$|\rho(x_1, x_3) - \rho(x_2, x_4)| < \varepsilon \,. \tag{D.2.26}$$

By compactness of $\mathcal{K}$, there exists a finite number of points $x_i \in \mathcal{K}$, $i \in I$ such that

$$\mathcal{K} \subset \bigcup_{i \in I} \mathbb{B}(x_i, \delta) \,. \tag{D.2.27}$$

We first show that the result holds for the points x_i before explaining why it actually suffices for the general case.

Fix $i, j \in I$. Since $x_i \sim x_j$, $B(x_i, x_j) = 0$ and therefore there exists sequences $\xi^k \in (\mathcal{X})^{N_k}$ with $\xi^k_0 = x_i$, $\xi^k_{N_k - 1} = x_j$ such that $\mathcal{A}_{N_k}(\xi^k) \to 0$ as $k \to \infty$. For the sake of contradiction, assume that from some k, there always exists $0 \leq n < N_k$ such that $d(\xi^k_n, \mathcal{K}) \geq \varepsilon$. Define n_k as the smallest n such that it happens, which is necessarily greater or equal to 1. Note that, by definition, $\xi^k_{n_k - 1}$ satisfies $d(\xi^k_{n_k - 1}, \mathcal{K}) \leq \varepsilon$.

Define $\mathcal{K}' \coloneqq \{x \in \mathcal{X} : d(x, \mathcal{K}) \leq \varepsilon\}$ which is compact. However, for k large enough, $\rho(\xi^k_{n_k - 1}, \xi^k_{n_k}) \leq \mathcal{A}_{N_k}(\xi^k) \leq 1$ so that $(\xi^k_{n_k - 1}, \xi^k_{n_k})$ belongs to $\Gamma_2^{\mathcal{K}'}(1)$, which is compact by Corollary C.2. Therefore, one can extract a subsequence from $(\xi^k_{n_k - 1}, \xi^k_{n_k},)_{k \geq 1}$ that converges to some $(x, z) \in \mathcal{X}$ that satisfies $d(z, \mathcal{K}) \geq \varepsilon$. Moreover, by l.s.c. of ρ, one has that

$$\begin{aligned} \rho(x, z) &\leq \liminf_{k \to \infty} \rho(\xi^k_{n_k - 1}, \xi^k_{n_k}) \\ &\leq \liminf_{k \to \infty} \mathcal{A}_{N_k}(\xi^k) = 0 \,, \end{aligned} \tag{D.2.28}$$

so that $\rho(x, z) = 0$. We now show that $x = z$ and that it is a critical point. By Lemma D.6, we have that

$$U_\infty(\xi^k_{n_k - 1}) - U_\infty(x_i) \leq 2B(x_i, \xi^k_{n_k - 1}) \leq 2\mathcal{A}_{N_k}(\xi^k) \tag{D.2.29a}$$

$$U_\infty(x_j) - U_\infty(\xi^k_{n_k}) \leq 2B(x_j, \xi^k_{n_k}) \leq 2\mathcal{A}_{N_k}(\xi^k) \,, \tag{D.2.29b}$$

so, taking the limit $k \to \infty$ yields

$$U_\infty(x) - U_\infty(x_i) \leq 0 \tag{D.2.30a}$$

$$U_\infty(x_j) - U_\infty(z) \leq 0 \,. \tag{D.2.30b}$$

However, the fact that x_i and x_j are equivalent and Lemma D.6 imply that $U_\infty(x_i) = U_\infty(x_j)$ so that $U_\infty(x) \leq U_\infty(z)$. Since α_∞ is increasing, we have that $f(x) \leq f(z)$.

But we have that $\rho(x, z) = 0$ so that $z = \Theta_1(x)$. Therefore, if $\nabla f(x) \neq 0$, we would have

$$f(z) - f(x) = -\int_0^1 \|\nabla f(\Theta_t(x))\|^2 \, dt < 0 \,, \tag{D.2.31}$$

which would be a contradiction. Therefore, $\nabla f(x) = 0$ and $x = z$. Since x is a critical point, to show that it belongs to $\mathcal{K}$, it suffices to show that $x_i \sim x$.

Take $\delta > 0$. By Lemma D.8, for δ small enough, $\mathcal{W}_\delta(\{x\})$ is an open neighborhood of x. Since $\xi^k_{n_k}$ converges to x, for k large enough, $\xi^k_{n_k}$ belongs to $\mathcal{W}_\delta(\{x\})$. $(\xi^k_0, \dots, \xi^k_{n_k}, x)$ and

$(x, \xi^k_{n_k}, \dots, \xi^k_{N_k-1})$ are, respectively, paths from x_i to x and from x to x_j with action cost of at most $\mathcal{A}_{N_k}(\xi^k) + \varepsilon$ so that, for k large enough, $B(x_i, x) \leq 2\varepsilon$ and $B(x, x_j) \leq 2\varepsilon$.

Hence, we have shown that $B(x_i, x) = B(x, x_j) = 0$. Then, we also have

$$0 \leq B(x, x_i) \leq B(x, x_j) + B(x_j, x_i) = 0\,, \tag{D.2.32}$$

and therefore $x \sim x_i$ and thus x belongs to $\mathcal{K}$. This is a contradiction.

Therefore, there must exist some sequence $\xi^{i,j} \in (\mathcal{X})^{N_{i,j}}$ with $\xi^{i,j}_0 = x_i$, $\xi^{i,j}_{N_{i,j}-1} = x_j$ such that $\mathcal{A}_{N_{i,j}}(\xi^{i,j}) < \varepsilon$ and $\max_{0 \leq n < N_{i,j}} d(\xi^{i,j}_n, \mathcal{K}) < \varepsilon$.

Finally, for the general class, consider $x, z \in \mathcal{K}$. By Eq. (D.2.27), there exists $i, j \in I$ such that $x \in \mathbb{B}(x_i, \delta)$, $z \in \mathbb{B}(x_j, \delta)$. Consider $\xi \in (\mathcal{X})^{N_{i,j}}$ a modification of $\xi^{i,j}$ defined by

$$\xi_n = \begin{cases} x & \text{if } n = 0, \\ \xi^{i,j}_n & \text{if } 0 < n < N_{i,j} - 1, \\ z & \text{if } n = N_{i,j} - 1, \end{cases} \tag{D.2.33}$$

that still satisfies $\max_{0 \leq n < N_{i,j}} d(\xi_n, \mathcal{K}) < \varepsilon$. Then, by uniform continuity of ρ, one has that

$$\mathcal{A}_{N_{i,j}}(\xi) < \mathcal{A}_{N_{i,j}}(\xi^{i,j}) + 2\varepsilon \leq 3\varepsilon\,, \tag{D.2.34}$$

which concludes the proof. ∎

The following lemma is inspired by Alongi & Nelson [1, Prop. 3.3.11].

Lemma D.10. *Equivalence classes are connected.*

Proof. Fix $\mathcal{K}$ an equivalence class of $\sim$.

For the sake of contradiction, assume that there are $\mathcal{U}$, $\mathcal{V}$ disjoint open sets of $\mathcal{X}$ such that both $\mathcal{U} \cap \mathcal{K}$ and $\mathcal{V} \cap \mathcal{K}$ are non-empty and $\mathcal{K} = (\mathcal{U} \cup \mathcal{V}) \cap \mathcal{K}$.

Take $x \in \mathcal{U} \cap \mathcal{K}$ and $x' \in \mathcal{V} \cap \mathcal{K}$. By Lemma D.9, there exists a sequence of paths $\xi^k \in (\mathcal{X})^{N_k}$ with $\xi^k_0 = x$, $\xi^k_{N_k-1} = x'$ for $k \geq 0$ such that $\mathcal{A}_{N_k}(\xi^k) \to 0$ as $k \to \infty$ and $\max_{0 \leq n < N_k} d(\xi^k_n, \mathcal{K}) \to 0$ as $k \to \infty$. Define a_k as the last point of ξ^k that belongs to $\mathcal{U}$ and b_k as the successor of a_k in ξ^k. Formally, a_k and b_k are defined by $a_k = \xi^k_{i_k}$ and $b_k = \xi^k_{i_k+1}$ where $i_k = \max\{i < N_k : \xi^k_i \in \mathcal{U}\}$.

By construction, both $d(a_k, \mathcal{K})$ and $d(b_k, \mathcal{K})$ go to zero as $k \to \infty$. In particular, both sequences lie in $\mathcal{U}_1(\mathcal{K})$ from some point onward, which is relatively compact, so they admit convergent subsequences. Without loss of generality, thus assume that $a_k \to a$ and $b_k \to b$ as $k \to \infty$ and a, b belong to $\mathcal{K}$.

Since a_k belongs to $\mathcal{U}$ for all $k \geq 0$, a_k is never in $\mathcal{V}$ so that a does not belong to $\mathcal{V}$ either. Since a belongs to $\mathcal{K}$, it must thus belong to $\mathcal{U}$. Similarly, b must belong to $\mathcal{V}$.

However, by construction, we have that $\rho(a_k, b_k) \leq \mathcal{A}_{N_k}(\xi^k) \to 0$ as $k \to \infty$ so that $\rho(a_k, b_k)$ converges to 0 as $k \to \infty$ as well. By l.s.c. of ρ (Corollary C.2 with $N = 1$), $\rho(a, b) = 0$ so that $b = F(a)$ by Lemma D.2. But since a belongs to $\mathcal{K}$, it is a critical point of f and therefore $a = F(a)$. Hence $a = b$ with $a \in \mathcal{U}$ and $b \in \mathcal{V}$, which is a contradiction. ∎

Lemma D.11. *Any connected component of* crit f *is included in a single equivalence class.*

Proof. Let $\mathcal{K}$ be a connected component of crit f and fix $x, x' \in \mathcal{K}$. We begin by considering a stronger version of Assumption 1(*c*), namely that there exists $\gamma \in \mathcal{C}([0,1], \mathcal{K})$ absolutely continuous such that $\gamma_0 = x$, $\gamma_1 = x'$, i.e., such that γ is differentiable almost everywhere with $\int_0^1 \|\dot\gamma_t\|\, dt < \infty$. We show that $x \sim x'$, i.e., that $B(x, x') = B(x', x) = 0$.

Let us begin by showing that $B(x, x') = 0$.

Since crit f is compact and $\mathcal{K}$ is closed as connected component of a closed set, $\mathcal{K}$ is compact. By invoking Lemma D.3 at every point of $\mathcal{K}$ and extracting a finite covering from

the family of balls obtained, we have that, there exists $\delta > 0$, $L > 0$ such that, for every $x \in \mathcal{K}$, $v \in \mathbb{B}(0, \delta)$,

$$\mathcal{L}(x, v) \leq L\|v\|^2 \,. \tag{D.2.35}$$

Fix $\varepsilon \in (0, \delta)$ and define, for any $t \in [0, 1]$,

$$\lambda_t \coloneqq \int_0^t \frac{\max(\|\dot{\gamma}_s\|, 1)}{\varepsilon}\, ds \,. \tag{D.2.36}$$

We have that λ is an increasing bijection from $[0, 1]$ to $[0, \lambda_1]$ and is absolutely continuous with $\dot{\lambda}_t = \max(\|\dot{\gamma}_t\|, 1)/\varepsilon$ almost everywhere. Consider $\tau \colon [0, \lambda_1] \to [0, 1]$ the inverse of λ, which is absolutely continuous with $\dot{\tau}_t = \varepsilon/\max(\|\dot{\gamma}_{\tau_t}\|, 1)$ almost everywhere. Define $\varphi \in \mathcal{C}([0, \lambda_1], \mathcal{X})$ by $\varphi_t = \gamma_{\tau_t}$ for any $t \in [0, \lambda_1]$. Then, φ is absolutely continuous and, for any $t \in [0, \lambda_1]$,

$$\dot{\varphi}_t = \dot{\gamma}_{\tau_t}\dot{\tau}_t = \frac{\varepsilon \dot{\gamma}_{\tau_t}}{\max(\|\dot{\gamma}_{\tau_t}\|, 1)} \tag{D.2.37}$$

which has norm less than $\varepsilon < \delta$.

Therefore, we have that,

$$\begin{aligned}
\mathcal{S}_{0,\lambda_1}(\varphi) &= \int_0^{\lambda_1} \mathcal{L}(\varphi_t, \dot{\varphi}_t)\, dt \\
&\leq L \int_0^{\lambda_1} \|\dot{\varphi}_t\|^2\, dt \\
&\leq \varepsilon L \int_0^{\lambda_1} \|\dot{\varphi}_t\|\, dt \\
&= \varepsilon L \int_0^{\lambda_1} \dot{\tau}_t \|\dot{\gamma}_{\tau_t}\|\, dt \\
&= \varepsilon L \int_0^{1} \|\dot{\gamma}_t\|\, dt
\end{aligned} \tag{D.2.38}$$

where the last equality is obtained by the change of variable $t \leftarrow \tau_t$. Thus, we have shown that, for any $\varepsilon \in (0, \delta)$,

$$B(x, x') \leq \varepsilon L \int_0^1 \|\dot{\gamma}_t\|\, dt \tag{D.2.39}$$

with $\int_0^1 \|\dot{\gamma}_t\|\, dt < \infty$ by construction so that $B(x, x') = 0$.

Reversing the roles of x and x' and considering the path $(\gamma_{1-t})_{t\in[0,1]}$ then yields that $B(x', x) = 0$. Therefore, we have shown that if there exists $\gamma \in \mathcal{C}([0, 1], \mathcal{K})$ absolutely continuous such that $\gamma_0 = x$, $\gamma_1 = x'$, then $x \sim x'$.

We now relax our assumption on the paths from absolute continuity to piecewise absolute continuity (Assumption 1*). For $x, x' \in \mathcal{K}$, by assumption, there exists $\gamma \in \mathcal{C}([0, 1], \mathcal{K})$ such that $\gamma_0 = x$, $\gamma_1 = x'$ and such that it is piecewise absolutely continuous: γ is differentiable almost everywhere and there exists $0 = t_0 < t_1 < \cdots < t_N = 1$ such that $\dot{\gamma}$ is integrable on every closed interval of (t_n, t_{n+1}) for $n = 0, \ldots, N-1$. Take $0 \leq n < N - 1$ and $t_n < s < t < t_{n+1}$. γ restricted to $[s, t]$ is absolutely continuous so that, by the previous case, all the points of $\{\gamma_s alt : u \in [s, t]\}$ are included in a single equivalence class $\mathcal{K}$. Taking $s \to t_n$ and $t \to t_{n+1}$ yields that $\{\gamma_s alt : u \in (t_n, t_{n+1})\}$ is included in $\mathcal{K}$. Moreover, by continuity of γ, γ_{t_n} and $\gamma_{t_{n+1}}$ belong to the closure of $\mathcal{K}$, which is closed by Lemma D.7, so that γ_{t_n} and $\gamma_{t_{n+1}}$ belong to $\mathcal{K}$ as well. Therefore, $\gamma_{t_n} \sim \gamma_{t_{n+1}}$. By transitivity, we obtain that $x = \gamma_0 \sim \gamma_{t_1} \sim \cdots \sim \gamma_{t_{N-1}} \sim \gamma_1 = x'$ so that $x \sim x'$. ∎

Combining Lemmas D.10 and D.11, we have shown that any connected component of crit f is included in a single equivalence class and since they are connected, two distinct connected components of crit f cannot belong to the same equivalent class; hence, we have that they coincide.

Corollary D.1. *Under the assumptions of Lemma D.11, the equivalence classes of $\sim$ are exactly connected components of* crit f.

We end this section by providing a sufficient condition for $B(x, x')$ to be finite.

Lemma D.12. *Consider $x, x' \in \mathcal{X}$ and assume that there exists $T > 0$, $\gamma \in \mathcal{C}^1([0,T], \mathcal{X})$ such that $\gamma_0 = x$, $\gamma_T = x'$ and such that, for every $t \in [0,T]$, $\nabla f(\gamma_t)$ is in the interior of the closed convex hull of the support of $\mathsf{U}(\gamma_t, \omega)$, i.e.,*

$$\nabla f(\gamma_t) \in \operatorname{int} \overline{\operatorname{conv}} \operatorname{supp} \mathsf{U}(\gamma_t, \omega)\,. \tag{D.2.40}$$

Then, $B(x, x') < \infty$.

Proof. By Brown [7, Thm. 3.6], Eq. (D.2.40) implies that, for every $t \in [0,T]$, $\nabla f(\gamma_t)$ belongs to $\nabla_p \bar{\mathcal{H}}(\gamma_t, \mathcal{X})$. Therefore, as in the proof of Lemma D.3, invoking the implicit function theorem on the equation

$$\nabla_p \bar{\mathcal{H}}(x, p) = \nabla f(x) + v \tag{D.2.41}$$

we obtain that there exists $\delta(\gamma_t) > 0$, $p\colon \mathbb{B}(\gamma_t, \delta(\gamma_t)) \times \mathbb{B}(0, \delta(\gamma_t)) \to \mathcal{X}$ such that,

$$\nabla_p \bar{\mathcal{H}}(x, p(x, v)) = \nabla f(x) + v\,, \tag{D.2.42}$$

or, equivalently

$$\nabla_p \mathcal{H}(x, p(x, v)) = v\,. \tag{D.2.43}$$

Therefore, as in the proof of Lemma D.3, we obtain that $\mathcal{L}$ is continuous on $\overline{\mathbb{B}}(\gamma_t, \delta(\gamma_t)/2) \times \overline{\mathbb{B}}(0, \delta(\gamma_t)/2)$ and therefore bounded by $M(\gamma_t) > 0$. Since γ is continuous, $\{\gamma_t : t \in [0,T]\}$ is compact and therefore, by extracting a finite covering from

$$\bigcup_{t \in [0,T]} \mathbb{B}(\gamma_t, \delta(\gamma_t)/2)\,, \tag{D.2.44}$$

we obtain that there exists $\delta > 0$ and $M > 0$ such that, for every $t \in [0,T]$, $\mathcal{L}(\gamma_t, \cdot)$ is finite and bounded by M on $\mathbb{B}(\gamma_t, \delta)$. Choosing a φ reparametrization of γ, which is $\mathcal{C}^1$, such that $\|\dot{\varphi}_t\| < \delta$ for every $t \in [0, S]$, we thus obtain a path such that

$$\mathcal{S}_{0,S}(\varphi) = \int_0^T alt\mathcal{L}(\varphi_t, \dot{\varphi}_t)\, dt \le MS < \infty\,, \tag{D.2.45}$$

which implies that $B(x, x') < \infty$. ∎

D.3. **Lyapunov condition.**

Definition 5 (Stopping times for the accelerated process)**.** For any set $S \subset \mathcal{X}$, we define the hitting and exit times of S:

$$\sigma_S \coloneqq \inf\{n \ge 1 : x_n^\eta \in S\} \tag{D.3.1a}$$

$$\tau_S \coloneqq \inf\{n \ge 0 : x_n^\eta \notin S\}\,. \tag{D.3.1b}$$

We will need the following concentration lemma and its corollary.

Lemma D.13 (Part of the proof of [60, Th. 1.19])**.** *Let X be a random variable in $\mathbb{R}^d$ such that, for all $v \in \mathbb{R}^d$,*

$$\log \mathbb{E}[\exp(\langle v, X\rangle)] \le \frac{\|v\|^2}{2}\,. \tag{D.3.2}$$

Then, for all $t > 0$, we have

$$\mathbb{P}\big(\|X\|^2 \geq t\big) \leq 6^d \exp\left(-\frac{t}{8}\right). \tag{D.3.3}$$

Corollary D.2. *In the context of Lemma D.13, it holds that*

$$\mathbb{E}\big[\|X\|^2\big] \leq 16d \log 6\,. \tag{D.3.4}$$

Proof. Since $\|X\|^2$ is non-negative, its expectation can be written as

$$\begin{aligned}\mathbb{E}\big[\|X\|^2\big] &= \int_0^{+\infty} \mathbb{P}\big(\|X\|^2 > t\big)\, dt \\ &\leq 8d \log 6 + \int_{8d\log 6}^{+\infty} \mathbb{P}\big(\|X\|^2 > t\big)\, dt\,. \end{aligned} \tag{D.3.5}$$

Invoking Lemma D.13 then yields that

$$\mathbb{E}\big[\|X\|^2\big] \leq 8d\log 6 + 6^d \int_{8d\log 6}^{+\infty} \exp\left(-\frac{t}{8}\right) dt \leq 8d \log 6 + 8\,, \tag{D.3.6}$$

which concludes the proof since $1 \leq d \log 6$. ∎

Lemma D.14 (Lyapunov condition)**.** *Define U_∞ as in Lemma D.6. Then, there exists $\mathcal{K} \subset \mathcal{X}$ compact, $\eta_0 > 0$, $c > 0$ such that, for any $\eta \leq \eta_0$, $n \geq 0$, if $x_n \notin \mathcal{K}$, then, almost surely,*

$$\begin{aligned} U_\infty(x_{n+1}) - U_\infty(x_n) &\leq \eta\left(\frac{\|\mathsf{U}(x_n,\omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))}\right) \\ &\leq \eta\left(\frac{\|\mathsf{U}(x_n,\omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - (16d\log 6 + c)\right). \end{aligned} \tag{D.3.7}$$

Proof. By Assumption 3**, there is $R \geq \frac{1}{2}$, $c > 0$ such that, for any $x \in \mathcal{X}$ such that $\|x\| \geq R$,

$$\begin{cases} f(x) \geq c \\ c \leq \frac{\sigma_\infty^2(f(x))}{\|x\|^s} \leq c^{-1} \\ \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \geq 16d\log 6 + c\,. \end{cases} \tag{D.3.8}$$

Then, define $\mathcal{K} \coloneqq \overline{\mathbb{B}}(0, 2R+1)$.

By definition, U_∞ is twice continuously differentiable and, its Hessian satisfies, for any $x \in \mathcal{X}$,

$$\operatorname{Hess} U_\infty(x) \preccurlyeq \frac{2\operatorname{Hess} f(x)}{\sigma_\infty^2(f(x))} \preccurlyeq \frac{2\beta(f)}{\sigma_\infty^2(f(x))} I\,. \tag{D.3.9}$$

For the sake of clarity, for any $n \geq 0$, denote by δx_n the quantity

$$\delta x_n = \frac{x_{n+1} - x_n}{\eta}\,. \tag{D.3.10}$$

For any $n \geq 0$, we now have

$$\begin{aligned} U_\infty(x_{n+1}) - U_\infty(x_n) &\leq \eta \langle \nabla U_\infty(x_n), \delta x_n\rangle \\ &\quad + \frac{\eta^2}{2}\|\delta x_n\|^2 \sup_{t\in[0,1]} \frac{2\beta(f)}{\sigma_\infty^2(f(x_n + t(x_{n+1} - x_n)))}\,. \end{aligned} \tag{D.3.11}$$

We first focus on bounding the last term. First note that, by the blanket assumptions, $\|x_{n+1} - x_n\| \leq 2\eta M(1 + \|x_n\|)$ so that, For any $t \in [0,1]$, $\eta \leq (4M)^{-1}$,

$$\|x_n + t(x_{n+1} - x_n)\| \geq \|x_n\| - \|x_{n+1} - x_n\|$$

$$\geq \|x_n\| - 2\eta M(1 + \|x_n\|)$$
$$\geq \frac{1}{2}(\|x_n\| - 1)\,. \tag{D.3.12}$$

If x_n is outside of $\mathcal{K}$, we have that $\|x_n\| \geq 2R+1$ and thus $\|x_n + t(x_{n+1} - x_n)\| \geq R$. Moreover, since $R \geq \frac{1}{2}$, x_n being outside of $\mathcal{K}$ also implies that $\frac{1}{2}\|x_n\| \geq 1$ and so $\|x_n + t(x_{n+1} - x_n)\| \geq \frac{1}{4}\|x_n\|$. By the definition of R, we thus have

$$\begin{aligned}\sigma_\infty^2(f(x_n + t(x_{n+1} - x_n))) &\geq c\|x_n + t(x_{n+1} - x_n)\|^s \\ &\geq c\left(\frac{1}{4}\right)^s \|x_n\|^s \\ &\geq \frac{c^2}{4^s}\sigma_\infty^2(f(x_n))\,. \end{aligned} \tag{D.3.13}$$

Thus, if $x_n \notin \mathcal{K}$, Eq. (D.3.9) yields that

$$U_\infty(x_{n+1}) - U_\infty(x_n) \leq \eta\langle \nabla U_\infty(x_n), \delta x_n\rangle + \eta^2 \frac{4^s\beta(f)}{c^2\sigma_\infty^2(f(x_n))}\|\delta x_n\|^2\,. \tag{D.3.14}$$

Now, we can rewrite the inner product as

$$\begin{aligned}\langle \nabla U_\infty(x_n), \delta x_n\rangle &= \frac{2\langle \nabla f(x_n), \delta x_n\rangle}{\sigma_\infty^2(f(x_n))} \\ &= \frac{\|\delta x_n + \nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\delta x_n\|^2}{\sigma_\infty^2(f(x_n))}\,. \end{aligned} \tag{D.3.15}$$

Plugging this into Eq. (D.3.14) and assuming that $\eta \leq \frac{c^2}{4^s\beta(f)}$, we obtain

$$\begin{aligned}U_\infty(x_{n+1}) - U_\infty(x_n) &\leq \eta\left(\frac{\|\delta x_n + \nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))}\right) \\ &\leq \eta\left(\frac{\|\mathsf{U}(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))}\right) \end{aligned} \tag{D.3.16}$$

where we unrolled δx_n in the last inequality. Using again that $x_n \notin \mathcal{K}$, we obtain

$$U_\infty(x_{n+1}) - U_\infty(x_n) \leq \eta\left(\frac{\|\mathsf{U}(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - (16d\log 6 + c)\right) \tag{D.3.17}$$

which concludes the proof. ∎

We will reuse, in later sections, the following fact that we thus state as a lemma: the sequence of iterates of SGD is *(weak) Feller* (see e.g., [24, Def. 4.4.2]).

Lemma D.15. *The Markov chain $(x_n)_{n\geq 0}$ is (weak) Feller.*

Proof. since both ∇f and U are continuous, for any $g\colon \mathcal{X} \to \mathbb{R}$ continuous and bounded, the function.

$$x \in \mathcal{X} \longmapsto \mathbb{E}_x[g(x_1)] = \mathbb{E}_x[g(x - \eta\nabla f(x) + \eta\mathsf{U}(x, \omega))] \tag{D.3.18}$$

is still continuous and bounded. Therefore, the Markov chain $(x_n)_{n\geq 0}$ is weak-Feller. ∎

Lemma D.16. *There is $\eta_0 > 0$ such that, for any $\eta \leq \eta_0$, there exists an invariant probability measure for $(x_n)_{n\geq 0}$.*

Proof. We invoke a general result on weak-Feller Markov chains that satisfy a Lyapunov condition, e.g., Hernández-Lerma & Lasserre [24, Thm. 7.2.4] or Douc et al. [14, Thm. 12.3.6].

First, Lemma D.15 ensures that $(x_n)_{n\geq 0}$ is weak-Feller.

Moreover, by Lemma D.14, there exists $\mathcal{K} \subset \mathcal{X}$ compact, $\eta_0 > 0$, $c > 0$ such that, for any $\eta \leq \eta_0$, $x_0 = x \notin \mathcal{K}$,

$$U_\infty(x_1) - U_\infty(x) \leq \eta \left(\frac{\|\mathsf{U}(x, \omega_0)\|^2}{\sigma_\infty^2(f(x))} - (16 d \log 6 + c) \right) . \tag{D.3.19}$$

Passing to the expectation yields that, for any $x \notin \mathcal{K}$,

$$\mathbb{E}_x[U_\infty(x_1)] - U_\infty(x) \leq \eta \left(\frac{\mathbb{E}_x\left[\|\mathsf{U}(x, \omega_0)\|^2\right]}{\sigma_\infty^2(f(x))} - (16 d \log 6 + c) \right) . \tag{D.3.20}$$

Appplying Corollary D.2 with $X \leftarrow \frac{\mathsf{U}(x,\omega_0)}{\sqrt{\sigma_\infty^2(f(x))}}$ (the conditions of application are verified from Assumption 2$^*(c)$) yields that

$$\mathbb{E}_x[U_\infty(x_1)] - U_\infty(x) \leq -\eta c . \tag{D.3.21}$$

Hence, for any $x \in \mathcal{X}$, it holds

$$\mathbb{E}_x[U_\infty(x_1)] - U_\infty(x) \leq -\eta c + \mathbb{1}\{x \in \mathcal{K}\} \left(\sup_{x' \in \mathcal{K}} \mathbb{E}_{x'}[U_\infty(x')] - \inf_{\mathcal{X}} U_\infty + \eta c \right) , \tag{D.3.22}$$

with U_∞ which is not identically equal to its minimum since f is coercive.

Therefore, we can apply Hernández-Lerma & Lasserre [24, Thm. 7.2.4] with $U_\infty - \inf_{\mathcal{X}} U_\infty$, that guarantees that there exists an invariant probability measure for $(x_n)_{n \geq 0}$. ∎

Lemma D.17. *There exists a compact set $\mathcal{D} \subset \mathcal{X}$, $\eta_0 > 0$, such that for any compact set $\mathcal{D}' \subset \mathcal{X}$ such that $\mathcal{D} \subset \mathcal{D}'$, there exists $a, b > 0$, such that*

$$\forall \eta \leq \eta_0 \, , \, x \in \mathcal{D}' \setminus \mathcal{D} \, , \, n \geq 0 \, , \, \mathbb{P}_x(\sigma_{\mathcal{D}} > n) \leq \exp\left(-\frac{an}{\eta} + \frac{b}{\eta} \right) . \tag{D.3.23}$$

Proof. This result is a consequence of Lemma D.14: there exists $\mathcal{K} \subset \mathcal{X}$ compact, $\eta_0 > 0$, $c > 0$ such that, for any $\eta \leq \eta_0$, $n \geq 0$, if $x_n \notin \mathcal{K}$, then, almost surely,

$$U_\infty(x_{n+1}) - U_\infty(x_n) \leq \eta \left(\frac{\|\mathsf{U}(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - (16 d \log 6 + c) \right) . \tag{D.3.24}$$

First, let us choose $\mathcal{D}$. There is some $R > 0$ such that $\mathcal{K} \subset \mathbb{B}(0, R)$. Define $\widetilde{R} \coloneqq e^{4M+1}(R+1)$ and $\mathcal{D} \coloneqq \overline{\mathbb{B}}(0, \widetilde{R})$. With $\eta \leq (4M)^{-1}$, if, for some $n \geq 0$, x_n is not in $\mathcal{D}$, by Lemma B.3, for any $k \leq \lceil \eta^{-1} \rceil$, x_{n+k} has norm greater or equal to R and thus x_{n+k} is not in $\mathcal{K}$ either.

Now, fix $N \geq 0$ and consider the event $\{\sigma_{\mathcal{D}} > N\}$ with $x_0 = x \in \mathcal{D}' \setminus \mathcal{D}$. This means that, for any $0 \leq n \leq N$, x_n^η is outside of $\mathcal{D}$ and so, for any $0 \leq n \leq N \lfloor \eta^{-1} \rfloor$, x_n is not in $\mathcal{K}$.

Summing Eq. (D.3.24) over $n = 0, \dots, N\lfloor \eta^{-1} \rfloor - 1$ yields

$$\begin{aligned} U_\infty(x_N^\eta) - U_\infty(x_0) &= U_\infty(x_{N\lfloor \eta^{-1} \rfloor}) - U_\infty(x_0) \\ &\leq \eta \sum_{n=0}^{N\lfloor \eta^{-1} \rfloor - 1} \left(\frac{\|\mathsf{U}(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - (16 d \log 6 + c) \right) . \end{aligned} \tag{D.3.25}$$

Define $\Delta \coloneqq \sup_{\mathcal{D}' \setminus \mathcal{D}} U_\infty - \inf_{\mathcal{X} \setminus \mathcal{D}} U_\infty$ which is finite since f is coercive. By definition, we have that $U_\infty(x_N^\eta) - U_\infty(x_0) \geq -\Delta$.

Therefore, on the event $\{\sigma_{\mathcal{D}} > N\}$, we have

$$\sum_{n=0}^{N\lfloor \eta^{-1} \rfloor - 1} \frac{\|\mathsf{U}(x_n, \omega_n)\|^2}{\sigma_\infty^2 \circ f(x_n)} \geq N \lfloor \eta^{-1} \rfloor (16 d \log 6 + c) - \frac{\Delta}{\eta} . \tag{D.3.26}$$

Now, on the whole event, consider the random variable $X \in \mathbb{R}^{N\lfloor\eta^{-1}\rfloor d}$ defined by

$$(X_{nd}+1,\dots,X_{(n+1)d}) = \frac{\mathsf{U}(x_n,\omega_n)}{\sqrt{\sigma_\infty^2 \circ f(x_n)}} \quad \text{for } n = 0,\dots,N\lfloor\eta^{-1}\rfloor - 1\,, \tag{D.3.27}$$

we have that $\mathbb{P}_x(\sigma_{\mathcal{D}} > N) \leq \mathbb{P}_x\Big(\|X\|^2 \geq N\lfloor\eta^{-1}\rfloor(16d\log 6 + c) - \frac{\Delta}{\eta}\Big)$.

Since, for any $v \in \mathbb{R}^d$,

$$\log \mathbb{E}\left[\exp\left(\left\langle v, \frac{\mathsf{U}(x_n,\omega_n)}{\sqrt{\sigma_\infty^2 \circ f(x_n)}}\right\rangle\right) \middle| x_0, x_1, \dots, x_n\right] \leq \frac{\|v\|^2}{2}\,, \tag{D.3.28}$$

the random variable X satisfies the assumptions of Lemma D.13 with $d \leftarrow N\lfloor\eta^{-1}\rfloor d$.

First, suppose that $t \coloneqq N\lfloor\eta^{-1}\rfloor(16d\log 6 + c) - \frac{\Delta}{\eta}$ is non-negative. Applying Lemma D.13 with this t yields that

$$\begin{aligned}
\mathbb{P}_x(\sigma_{\mathcal{D}} > N) &\leq \mathbb{P}_x\left(\|X\|^2 \geq N\lfloor\eta^{-1}\rfloor(16d\log 6 + c) - \frac{\Delta}{\eta}\right) \\
&\leq 6^{N\lfloor\eta^{-1}\rfloor d}\exp\left(-\frac{N\lfloor\eta^{-1}\rfloor(16d\log 6 + c) - \frac{\Delta}{\eta}}{8}\right) \\
&= \exp\left(-\frac{N\lfloor\eta^{-1}\rfloor(8d\log 6 + c) - \frac{\Delta}{\eta}}{8}\right).
\end{aligned} \tag{D.3.29}$$

If t, defined above, is negative, then in particular $\frac{\Delta}{\eta} \geq N\lfloor\eta^{-1}\rfloor c$ so that this bounds still (trivially) holds.

Finally, in particular, for $\eta \leq 1/2$, $\lfloor\eta^{-1}\rfloor \geq (2\eta)^{-1}$ so we obtain

$$\mathbb{P}_x(\sigma_{\mathcal{D}} > N) \leq \exp\left(-\frac{N(8d\log 6 + c)}{16\eta} + \frac{\Delta}{8\eta}\right) \tag{D.3.30}$$

and our proof is complete. ∎

D.4. **Preliminary estimates and lemmas.** We will use the following lemma, which corresponds to Kifer [35, Lem. 5.3].

Lemma D.18. *Let $\mathcal{K} \subset \mathcal{X}$ be a compact set such that $\mathcal{K} \cap \operatorname{crit} f = \emptyset$. Then there exists $c > 0$, $N \geq 1$, $\eta_0 > 0$, such that, for any $n > N$, $x \in \mathcal{K}$, $\eta \leq \eta_0$,*

$$\mathbb{P}_x\big(\sigma_{\mathcal{X}\setminus\mathcal{K}} > n\big) = \mathbb{P}_x(\tau_{\mathcal{K}} > n) \leq \exp(-c(n-N)/\eta)\,. \tag{D.4.1}$$

Proof. The proof is exactly the same as the proof Kifer [35, Lem. 5.3], which only uses the l.s.c. of $\mathcal{A}_N$ (Corollary C.2). ∎

The following lemma provides a convenient reformulation of the results of Lemma D.17 and Lemma D.18.

Lemma D.19.

- *There exists $\mathcal{D} \subset \mathcal{X}$ a compact set, $\eta_0 > 0$, such that for any $\mathcal{D}' \subset \mathcal{X}$ compact set such that $\mathcal{D} \subset \mathcal{D}'$, there exists $\alpha_0, a, b > 0$ such that,*

$$\forall \eta \leq \eta_0,\ \alpha \leq \alpha_0,\ x \in \mathcal{D}' \setminus \mathcal{D}\,, \quad \mathbb{E}_x\left[e^{\frac{\alpha\sigma_{\mathcal{D}}}{\eta}}\right] \leq e^{\frac{a\alpha}{\eta}+b}\,. \tag{D.4.2}$$

- *For any $\mathcal{K} \subset \mathcal{X}$ compact such that $\mathcal{K} \cap \operatorname{crit} f = \emptyset$, there exists $\eta_0, \alpha_0, a, b > 0$ such that,*

$$\forall \eta \leq \eta_0,\ \alpha \leq \alpha_0,\ x \in \mathcal{K}\,, \quad \mathbb{E}_x\left[e^{\frac{\alpha\tau_{\mathcal{K}}}{\eta}}\right] \leq e^{\frac{a\alpha}{\eta}+b}\,. \tag{D.4.3}$$

Proof. The proofs of both statements are very similar, so we prove only the second one for notational convenience. Fix $\mathcal{K} \subset \mathcal{X}$ compact such that $\mathcal{K} \cap \operatorname{crit} f = \emptyset$. By Lemma D.18, there exists $c > 0$, $N \geq 1$ such that, for any $n > N$, $x \in \mathcal{K}$, $\eta \leq \eta_0$,

$$\mathbb{P}_x(\tau_{\mathcal{K}} > n) \leq \exp(-c(n-N)/\eta)\,. \tag{D.4.4}$$

Let us first bound $\mathbb{E}_x\left[\exp\left(\frac{\alpha(\tau_{\mathcal{K}}-N)}{\eta}\right)\right]$. We have that

$$\begin{aligned}
\mathbb{E}_x\left[\exp\left(\frac{\alpha(\tau_{\mathcal{K}}-N)}{\eta}\right)\right] &= \int_0^\infty \mathbb{P}_x\left(\exp\left(\frac{\alpha(\tau_{\mathcal{K}}-N)}{\eta}\right) > t\right) dt \\
&\leq e^{\frac{\alpha}{\eta}} + \int_{e^{\frac{\alpha}{\eta}}}^\infty \mathbb{P}_x\left(\tau_{\mathcal{K}} > N + \frac{\eta \log t}{\alpha}\right) dt \\
&\leq e^{\frac{\alpha}{\eta}} + \int_{e^{\frac{\alpha}{\eta}}}^\infty \mathbb{P}_x\left(\tau_{\mathcal{K}} > N + \left\lfloor \frac{\eta \log t}{\alpha} \right\rfloor\right) dt \\
&\leq e^{\frac{\alpha}{\eta}} + \int_{e^{\frac{\alpha}{\eta}}}^\infty \exp\left(-\frac{c}{\eta}\left\lfloor \frac{\eta \log t}{\alpha} \right\rfloor\right) dt\,,
\end{aligned} \tag{D.4.5}$$

where we used Eq. (D.4.4) in the last inequality.

Lower bounding $\lfloor s \rfloor$ by $s-1$, we obtain that

$$\begin{aligned}
\mathbb{E}_x\left[\exp\left(\frac{\alpha(\tau_{\mathcal{K}}-N)}{\eta}\right)\right] &\leq e^{\frac{\alpha}{\eta}} + \int_{e^{\frac{\alpha}{\eta}}}^\infty \exp\left(-\frac{c}{\eta}\frac{\eta \log t}{\alpha} + \frac{c}{\eta}\right) dt \\
&= e^{\frac{\alpha}{\eta}} + \int_{e^{\frac{\alpha}{\eta}}}^\infty e^{\frac{c}{\eta}} t^{-\frac{c}{\alpha}}\, dt\,.
\end{aligned} \tag{D.4.6}$$

Performing the change of variable $s \leftarrow e^{-\frac{\alpha}{\eta}} t$, we obtain that

$$\mathbb{E}_x\left[\exp\left(\frac{\alpha(\tau_{\mathcal{K}}-N)}{\eta}\right)\right] \leq e^{\frac{\alpha}{\eta}}\left(1 + \int_1^\infty s^{-\frac{c}{\alpha}}\, ds\right). \tag{D.4.7}$$

When $\alpha \leq c/2$, we obtain that

$$\mathbb{E}_x\left[\exp\left(\frac{\alpha(\tau_{\mathcal{K}}-N)}{\eta}\right)\right] \leq e^{\frac{\alpha}{\eta}}\left(1 + \int_1^\infty s^{-2}\, ds\right), \tag{D.4.8}$$

and therefore,

$$\mathbb{E}_x\left[\exp\left(\frac{\alpha\tau_{\mathcal{K}}}{\eta}\right)\right] \leq e^{\frac{(N+1)\alpha}{\eta}}\left(1 + \int_1^\infty s^{-2}\, ds\right), \tag{D.4.9}$$

which concludes the proof. ∎

The following lemma upper-bounds the probability of exiting a large neighborhood of the critical points before visiting a smaller one critical points.

Lemma D.20. *Consider* $\operatorname{crit} f \subset \mathcal{U} \subset \mathcal{D} \subset \mathcal{X}$ *with* $\mathcal{U}$ *an open set and* $\mathcal{D}$ *a compact set. There exists* $\mathcal{D}' \subset \mathcal{X}$ *compact set such that* $\mathcal{D} \subset \mathcal{D}'$, $\Delta > 0$, $\eta_0 > 0$ *such that, for any* $\eta \leq \eta_0$, $x \in \mathcal{D}$,

$$\mathbb{P}_x\left(\tau_{\mathcal{D}'} < \sigma_{\mathcal{U}}\right) \leq \exp\left(-\frac{\Delta}{\eta}\right). \tag{D.4.10}$$

Proof. Define $\mathcal{D}' \coloneqq \{x \in \mathcal{X} : f(x) \leq \sup_{\mathcal{D}} f + 1\}$ and let U_∞ be as in Lemma D.6.

Since α_∞ is (stricly) increasing as its derivative is (stricly) positive by definition, we have

$$\Delta \coloneqq \alpha_\infty\left(\sup_{\mathcal{D}} f + 1\right) - \alpha_\infty\left(\sup_{\mathcal{D}} f\right) > 0\,. \tag{D.4.11}$$

By Lemma D.18 applied to $\mathcal{K} \leftarrow \mathcal{D}'_\delta \setminus \mathcal{U}$, there exists $c > 0$, $N_0 \geq 1$, $\eta_0 > 0$ such that for any $n > N_0$, $x \in \mathcal{D}'_\delta \setminus \mathcal{U}$, $\eta \leq \eta_0$,

$$\mathbb{P}_x\Big(\tau_{\mathcal{D}'_\delta \setminus \mathcal{U}} > n\Big) \leq \exp(-c(n - N_0)/\eta)\,. \tag{D.4.12}$$

Defining $N \coloneqq \lceil \frac{\Delta}{c} \rceil + N_0$, which is greater or equal than 1, we obtain that, for any $\eta \leq \eta_0$, $x \in \mathcal{D}'_\delta \setminus \mathcal{U}$,

$$\mathbb{P}_x\Big(\tau_{\mathcal{D}'_\delta \setminus \mathcal{U}} \geq N\Big) \leq \exp\left(-\frac{\Delta}{\eta}\right). \tag{D.4.13}$$

Note that this inequality actually holds for any $x \in \mathcal{D}'_\delta$.

We now bound $\mathbb{P}_x\Big(\tau_{\mathcal{D}'_\delta} > \sigma_{\mathcal{U}}\Big)$ for $x \in \mathcal{D}$ by distinguishing the cases where $\tau_{\mathcal{D}'_\delta} < N$ and $\tau_{\mathcal{D}'_\delta} \geq N$. For any $x \in \mathcal{D}$, we have that

$$\begin{aligned}
\mathbb{P}_x\Big(\tau_{\mathcal{D}'_\delta} < \sigma_{\mathcal{U}}\Big) &\leq \mathbb{P}_x\Big(\tau_{\mathcal{D}'_\delta} < \sigma_{\mathcal{U}},\, \tau_{\mathcal{D}'_\delta} < N\Big) + \mathbb{P}_x\Big(\tau_{\mathcal{D}'_\delta} < \sigma_{\mathcal{U}},\, \tau_{\mathcal{D}'_\delta} \geq N\Big) \\
&\leq \mathbb{P}_x\Big(\tau_{\mathcal{D}'_\delta} < N\Big) + \mathbb{P}_x\Big(\tau_{\mathcal{D}'_\delta \setminus \mathcal{U}} \geq N\Big) \\
&\leq \mathbb{P}_x\Big(\tau_{\mathcal{D}'_\delta} < N\Big) + \exp\left(-\frac{\Delta}{\eta}\right)
\end{aligned} \tag{D.4.14}$$

where we used Eq. (D.4.13).

We now focus on bounding the first term. For this, we first show that $\tau_{\mathcal{D}'_\delta} < N$ implies that

$$\operatorname{dist}_N\Big(x^\eta, \Gamma_N^{\{x\}}(\Delta/4)\Big) > \delta/2\,. \tag{D.4.15}$$

For the sake of contradiction, suppose that this inequality does not hold. Therefore, there must exist $\xi \in \Gamma_N^{\{x\}}(\Delta/4)$ such that $\operatorname{dist}_N(x^\eta, \xi) < \delta$. In particular, there is some $n < N$ such that $\xi_n \notin \mathcal{D}'$, so that, by Definition 3,

$$\mathcal{A}_N(\xi) \geq \mathcal{A}_{n+1}(\xi) \geq B(\xi_0, \xi_n)\,. \tag{D.4.16}$$

By Lemma D.6, we have that

$$\begin{aligned}
B(\xi_0, \xi_n) &\geq \frac{1}{2}(U_\infty(\xi_n) - U_\infty(\xi_0)) \\
&\geq \frac{1}{2}\left(\alpha_\infty\left(\inf_{\mathcal{X} \setminus \mathcal{D}'} f\right) - \alpha_\infty\left(\sup_{\mathcal{D}'} f\right)\right),
\end{aligned} \tag{D.4.17}$$

since α_∞ is increasing. By construction of $\mathcal{D}'$, we have further that

$$B(\xi_0, \xi_n) \geq \frac{1}{2}\left(\alpha_\infty\left(\sup_{\mathcal{D}} f + 1\right) - \alpha_\infty\left(\sup_{\mathcal{D}} f)\right)\right) = \frac{\Delta}{2}\,, \tag{D.4.18}$$

so that $\mathcal{A}_N(\xi) \geq \Delta/2$, which is a contradiction with $\xi \in \Gamma_N^{\{x\}}(\Delta/4)$.

Therefore, we have that

$$\begin{aligned}
\mathbb{P}\Big(\tau_{\mathcal{D}'_\delta} < N\Big) &\leq \mathbb{P}\Big(\operatorname{dist}_N\Big(x^\eta, \Gamma_N^{\{x\}}(\Delta/4)\Big) > \delta/2\Big) \\
&\leq \exp\left(-\frac{\Delta}{8\eta}\right)
\end{aligned} \tag{D.4.19}$$

where we invoked Corollary C.2 with $\delta \leftarrow \delta/2$, $s \leftarrow \Delta/2$, $\varepsilon \leftarrow \Delta/8$. ∎

The following lemma is key to our analysis: it shows that SGD spends most of its time near its critical points.

Lemma D.21. *Consider* crit $f \subset \mathcal{U} \subset \mathcal{D} \subset \mathcal{X}$ *with* $\mathcal{U}$ *an open set and* $\mathcal{D}$ *a compact set. Then, there is some* $\eta_0, \alpha_0, a, b > 0$ *such that,*

$$\forall \eta \le \eta_0,\, \alpha \le \alpha_0,\, x \in \mathcal{D}\,, \quad \mathbb{E}_x\Big[e^{\frac{\alpha\sigma_{\mathcal{U}}}{\eta}}\Big] \le e^{\frac{a\alpha}{\eta}+b}\,. \tag{D.4.20}$$

Proof. Fix $\varepsilon > 0$. Without loss of generality, assume that $\mathcal{D}$ is large enough to include the compact set given by the first item of Lemma D.19 (note that the guarantee of the first item of Lemma D.19 still holds even if $\mathcal{D}$ is larger).

Apply Lemma D.20 with $\mathcal{U} \leftarrow \mathcal{U}$, $\mathcal{D} \leftarrow \mathcal{D}$ and denote by $\tilde{\mathcal{D}}$ the obtained compact and $\eta_0, \Delta > 0$ such that, for every $\eta \le \eta_0$, $x \in \mathcal{D}$,

$$\mathbb{P}_x(\tau_{\tilde{\mathcal{D}}} < \sigma_{\mathcal{U}}) \le \exp\left(-\frac{\Delta}{\eta}\right). \tag{D.4.21}$$

Define $r \coloneqq \sup_{x\in\tilde{\mathcal{D}}}\|x\|$ and $R = e^{8M}(1+r)$. Assuming that $\eta \le 1$, by Lemma B.2, for any $x \in \tilde{\mathcal{D}}$, the next two iterates of $(x_n^\eta)_{n\ge0}$ satisfies $\|x_1^\eta\| \le R$ and $\|x_2^\eta\| \le R$. Define $\mathcal{D}' \coloneqq \overline{\mathbb{B}}(0,R)$.

We invoke both items of Lemma D.19 with $\mathcal{K} \leftarrow \tilde{\mathcal{D}} \setminus \mathcal{U}$ and denote by $c > 0$ a constant that satisfies the bounds of both items. In the rest of the proof, we consider η and α smaller than the bounds given by this lemma.

Our goal is to bound, for any $N \ge 0$, the quantity,

$$s_N(\alpha,\eta) \coloneqq \sup_{x\in\mathcal{D}} \mathbb{E}_x\left[\exp\left(\frac{\alpha\sigma_{\mathcal{U}}^N}{\eta}\right)\right], \quad \text{where} \quad \sigma_{\mathcal{U}}^N \coloneqq \min(N, \sigma_{\mathcal{U}})\,. \tag{D.4.22}$$

Note that, by construction, $s_N(\alpha,\eta)$ is finite.

Take $x \in \mathcal{D}$. In particular, Lemma D.19 implies that $\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}} < +\infty$ almost surely for all $x \in \mathcal{D}$.

$$\begin{aligned}
\exp\left(\frac{\alpha\sigma_{\mathcal{U}}^N}{\eta}\right) &= \mathbb{1}\{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}} \in \mathcal{U}\}\exp\left(\frac{\alpha\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}{\eta}\right) \\
&\qquad + \mathbb{1}\{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}} \notin \mathcal{D}_\delta\}\exp\left(\frac{\alpha\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}{\eta}\right)\exp\left(\frac{\alpha\left(\sigma_{\mathcal{U}}^N - \tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}\right)}{\eta}\right) \\
&\le \mathbb{1}\{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}} \in \mathcal{U}\}\exp\left(\frac{\alpha\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}{\eta}\right) \\
&\qquad + \mathbb{1}\{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}} \notin \mathcal{D}_\delta\}\exp\left(\frac{\alpha\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}{\eta}\right)\exp\left(\frac{\alpha\min\left(\sigma_{\mathcal{U}} - \tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}, N\right)}{\eta}\right)
\end{aligned} \tag{D.4.23}$$

so that we can apply the strong Markov property to the Markov chain $(x_n^\eta)_{n\ge0}$ with stopping time $\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}$ to obtain that

$$\begin{aligned}
\mathbb{E}_x\left[\exp\left(\frac{\alpha\sigma_{\mathcal{U}}^N}{\eta}\right)\right] \le \mathbb{E}_x\Bigg[&\mathbb{1}\{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}} \in \mathcal{U}\}\exp\left(\frac{\alpha\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}{\eta}\right) \\
&+ \mathbb{1}\{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}} \notin \mathcal{D}_\delta\}\exp\left(\frac{\alpha\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}{\eta}\right)\mathbb{E}_{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}}\left[\exp\left(\frac{\alpha\sigma_{\mathcal{U}}^N}{\eta}\right)\right]\Bigg]
\end{aligned} \tag{D.4.24}$$

We now bound

$$\mathbb{1}\{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}} \notin \mathcal{D}_\delta\}\,\mathbb{E}_{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}}\left[\exp\left(\frac{\alpha\sigma_{\mathcal{U}}^N}{\eta}\right)\right] = \mathbb{E}_{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}}\left[\mathbb{1}\{x_0^\eta \notin \mathcal{D}_\delta\}\exp\left(\frac{\alpha\sigma_{\mathcal{U}}^N}{\eta}\right)\right] \tag{D.4.25}$$

Since x is in $\mathcal{D}$, by definition, $\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}$ is at least equal to 1 and $x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}-1}$ is still in $\tilde{\mathcal{D}}$. By definition of $\mathcal{D}'$, $x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}$ must still be in $\mathcal{D}'$. Therefore, the guarantee of Lemma D.19 applies and, since in particular it implies that $\sigma_{\mathcal{D}}$ is finite almost surely when the chain is started at $x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}$, we can apply the strong Markov property to obtain that

$$\begin{aligned}\mathbb{E}_{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}}\left[\mathbb{1}\{x_0\notin\tilde{\mathcal{D}}\}\exp\left(\frac{\alpha\sigma^N_{\mathcal{U}}}{\eta}\right)\right] &\leq \mathbb{E}_{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}}\left[\mathbb{1}\{x_0\notin\tilde{\mathcal{D}}\}\exp\left(\frac{\alpha\sigma_{\mathcal{D}}}{\eta}\right)\mathbb{E}_{x^\eta_{\sigma_{\mathcal{D}}}}\left[\exp\left(\frac{\alpha\sigma^N_{\mathcal{U}}}{\eta}\right)\right]\right]\\ &\leq s_N(\alpha,\eta)\,\mathbb{E}_{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}}\left[\mathbb{1}\{x_0\notin\tilde{\mathcal{D}}\}\exp\left(\frac{\alpha\sigma_{\mathcal{D}}}{\eta}\right)\right] \qquad \text{(D.4.26)}\end{aligned}$$

where, for the second inequality, we used the definition of $s_N(\alpha,\eta)$ and the fact that $x_{\sigma_{\mathcal{D}}}$ is in $\mathcal{D}$.

Now, to bound the remaining expectation, note that x_0 does not belong to $\tilde{\mathcal{D}}$ and, *a fortiori*, does not belong to $\mathcal{D}$. Therefore, $\sigma_{\mathcal{D}}$ depends only on $(x^\eta_n)_{n\geq 1}$ and the (weak) Markov property implies that

$$\begin{aligned}\mathbb{E}_{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}}\left[\mathbb{1}\{x_0\notin\tilde{\mathcal{D}}\}\exp\left(\frac{\alpha\sigma_{\mathcal{D}}}{\eta}\right)\right] &\leq \mathbb{E}_{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}}\left[\mathbb{1}\{x_0\notin\tilde{\mathcal{D}}\}\,\mathbb{E}_{x_1}\left[\exp\left(\frac{\alpha(1+\sigma_{\mathcal{D}})}{\eta}\right)\right]\right]\\ &\leq \exp\left(\frac{\alpha(1+c)}{\eta}\right)\mathbb{E}_{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}}\left[\mathbb{1}\{x_0\notin\tilde{\mathcal{D}}\}\right], \qquad \text{(D.4.27)}\end{aligned}$$

by Lemma D.19, since the $x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}-1}$ is still in $\tilde{\mathcal{D}}$ so that x_1 of the chain started at $x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}$ is still in $\mathcal{D}'$.

Lemma D.20 then implies that,

$$\mathbb{E}_{x^\eta_{\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}}\left[\mathbb{1}\{x_0\notin\tilde{\mathcal{D}}\}\exp\left(\frac{\alpha\sigma_{\mathcal{D}'}}{\eta}\right)\right]\leq\exp\left(\frac{\alpha(1+a)-\Delta}{\eta}+b\right) \qquad \text{(D.4.28)}$$

Combining these bounds with Eq. (D.4.24) then gives

$$\begin{aligned}\mathbb{E}_x\left[\exp\left(\frac{\alpha\sigma^N_{\mathcal{U}}}{\eta}\right)\right] &\leq \left(1+s_N(\alpha,\eta)\exp\left(\frac{\alpha(1+a)-\Delta}{\eta}+b\right)\right)\mathbb{E}_x\left[\exp\left(\frac{\alpha\tau_{\tilde{\mathcal{D}}\setminus\mathcal{U}}}{\eta}\right)\right]\\ &\leq \left(1+s_N(\alpha,\eta)\exp\left(\frac{\alpha(1+a)-\Delta}{\eta}+b\right)\right)\exp\left(\frac{a\alpha}{\eta}+b\right), \qquad \text{(D.4.29)}\end{aligned}$$

by Lemma D.19. Since this inequality is valid for any $x\in\mathcal{D}$, we have shown that

$$s_N(\alpha,\eta)\leq\left(e^{\frac{a\alpha}{\eta}+b}+s_N(\alpha,\eta)e^{\frac{\alpha(1+2a)-\Delta}{\eta}+b}\right). \qquad \text{(D.4.30)}$$

For $\alpha\leq\Delta/(2(1+2a))$ and η small enough,

$$e^{\frac{\alpha(1+2a)-\Delta}{\eta}+b}\leq e^{-\frac{\Delta}{2\eta}+b}\leq\frac{1}{2}, \qquad \text{(D.4.31)}$$

and we obtain that

$$s_N(\alpha,\eta)\leq 2e^{\frac{a\alpha}{\eta}+b}. \qquad \text{(D.4.32)}$$

Taking $N\to+\infty$ and using Fatou's lemma yields that

$$\sup_{x\in\mathcal{D}}\mathbb{E}_x\left[\exp\left(\frac{\alpha\sigma_{\mathcal{U}}}{\eta}\right)\right]\leq 2e^{\frac{a\alpha}{\eta}+b}, \qquad \text{(D.4.33)}$$

which concludes the proof. ∎

D.5. **Estimates of the invariant measure.** Define, for sets $A, B \subset \mathcal{X}$,

$$B(A,B) \coloneqq \inf\{B(x,z) : x \in A, z \in B\}\,. \tag{D.5.1}$$

Recall that for any i, j, we define

$$B_{i,j} = B(\mathcal{K}_i, \mathcal{K}_j) = \inf\{B(x,z) : x \in \mathcal{K}_i, z \in \mathcal{K}_j\}\,. \tag{D.5.2}$$

Recall that, to avoid degenerate cases, we assume that Assumption 4 holds, that is,

$$B_{ij} < \infty \quad \text{for all } i, j = 1, \dots, K. \tag{D.5.3}$$

With Lemma D.12, this assumption is satisfied in particular if the following condition holds: for any $i, j = 1, \dots, K$, there exists a $\mathcal{C}^1$ path γ joining $\mathcal{K}_i$ and $\mathcal{K}_j$ such that, for all t, $\nabla f(\gamma_t)$ belongs to the interior of the closed convex hull of the support of the noise $\mathsf{U}(\gamma_t, \omega)$:

$$\nabla f(\gamma_t) \in \operatorname{int} \overline{\operatorname{conv}} \operatorname{supp} \mathsf{U}(\gamma_t, \omega)\,, \tag{D.5.4}$$

This means there is a sufficient level of noise for the probability of going from $\mathcal{K}_i$ to $\mathcal{K}_j$ with at least one path to be non-zero (though it can be vanishingly small). Note that it does not constrain the nature of the noise itself — which can be discrete, continuous or else —, only the support of its distribution.

Moreover, if Assumption 4 did not hold, the same analysis as in this section could still be carried out. We would consider the components of the graph $\mathcal{G}$ connected by edges with finite weights proceed with the proof on each of them, and obtain the same results on each of these components. To keep the complexity of the proof reasonable, we will not consider this case here.

We adapt to our context Kifer [35, Lem. 5.4] and simplify it using ideas from Freidlin & Wentzell [20, Chap. 6].

Definition 6 (Freidlin & Wentzell [20, Chap. 6,§2])**.** For i, j,

$$\widetilde{B}_{i,j} \coloneqq \inf\left\{ \mathcal{A}_N(\xi) : N \geq 1, \xi \in \mathcal{X}^N,\, \xi_0 \in \mathcal{K}_i,\, \xi_{N-1} \in \mathcal{K}_j,\, \xi_n \notin \bigcup_{l \neq i,j} \mathcal{K}_l \text{ for } n = 1, \dots, N-2 \right\}. \tag{D.5.5}$$

We now defined an important object: the law of the (accelerated) iterated at the first time they reach some set $\mathcal{V}$ (typically a neighborhood of the critical set), following e.g., Douc et al. [14, Chap. 3.4].

Definition 7 (Kifer [35, Prop. 5.3])**.** For $\mathcal{V}$ open set, with hitting time (of the accelerated sequence) $\sigma_{\mathcal{V}} \coloneqq \inf\{n \geq 1 : x_n^\eta \in \mathcal{V}\}$, (as in Definition 5), denote the law of $x_{\sigma_{\mathcal{V}}}^\eta$ started at x by $\mathbb{Q}_{\mathcal{V}}(x, \cdot)$ and the corresponding N-step transition probability by $\mathbb{Q}_{\mathcal{V}}^N(x, \cdot)$.

In words, $\mathbb{Q}_{\mathcal{V}}(x, \cdot)$ is the distributions of the x^η started at x at the first time they reach $\mathcal{V}$. Typically, $\mathbb{Q}_{\mathcal{V}}(x, \mathcal{U})$ is the probability that $\mathcal{U} \subset \mathcal{V}$ is reached first among $\mathcal{V}$. Then, $\mathbb{Q}_{\mathcal{V}}^N(x, \cdot)$ is the distribution of the x^η started at x at the N-th time they reach $\mathcal{V}$.

We first give estimates of the transition probabilities using the $\widetilde{B}_{i,j}$. We will then translate them to $B_{i,j}$.

Lemma D.22. *For any $\varepsilon > 0$, for any small enough neighborhoods $\mathcal{V}_i$ of $\mathcal{K}_i$, $i = 1, \dots, K$, there is some $\eta_0 > 0$ such that for all i, j, $x \in \mathcal{V}_i$, $0 < \eta < \eta_0$,*

$$\mathbb{Q}_{\mathcal{V}}(x, \mathcal{V}_j) \leq \exp\left(-\frac{\widetilde{B}_{i,j}}{\eta} + \frac{\varepsilon}{\eta} \right). \tag{D.5.6}$$

where we defined $\mathcal{V} \coloneqq \bigcup_{i=1}^K \mathcal{V}_i$.

Proof. Assume that, without loss of generality, ε is small enough so that Lemma D.8 with $r \leftarrow \varepsilon$ can be applied to every $\mathcal{K}_i$, $i = 1, \dots, K$. Denote by $\mathcal{W}_i$, $i = 1, \dots, K$ the corresponding neighborhoods of $\mathcal{K}_i$.

Since these $\mathcal{W}_i$'s are open neighborhoods of the $\mathcal{K}_i$'s, there exists $\delta > 0$ such that $\mathcal{U}_{2\delta}(\mathcal{K}_i) \subset \mathcal{W}_i$ for all $i = 1, \dots, K$. Require then that $\mathcal{V}_i$ be contained in $\mathcal{U}_\delta(\mathcal{K}_i)$ so that $\mathcal{U}_\delta(\mathcal{V}_i) \subset \mathcal{W}_i$ for all $i = 1, \dots, K$. Moreover, assume that $\delta > 0$ is small enough so that the neighborhoods $\mathcal{U}_\delta(\mathcal{K}_i)$, $i = 1, \dots, K$ are pairwise disjoint.

Define $0 < \delta' \leq \delta$ such that $\mathcal{U}'_\delta(\mathcal{K}_i)$ is contained in $\mathcal{V}_i$ for all $i = 1, \dots, K$.

Fix $i, j \in I$ and consider $\xi \in (\mathcal{X})^N$ such that $\xi_0 \in \mathcal{U}'_\delta(\mathcal{V}_i) \subset \mathcal{W}_i$, $\xi_{N-1} \in \mathcal{U}'_\delta(\mathcal{V}_j) \subset \mathcal{W}_j$ and $\xi_n \in \mathcal{U}_{\delta'}\Big(\mathcal{X} \setminus \bigcup_{l \neq i,j} \mathcal{V}_l\Big)$ for all $n = 1, \dots, N-2$. By the choice of δ', ξ_n cannot be in $\bigcup_{l \neq i,j} \mathcal{K}_l$ for any $n = 1, \dots, N-2$.

By definition of $\mathcal{W}_i$ and $\mathcal{W}_j$, there are $x \in \mathcal{K}_i$, $x' \in \mathcal{K}_j$ such that $\rho(x, \xi_0) < \varepsilon$, $\rho(\xi_{N-1}, x') < \varepsilon$. Therefore, the path $\zeta \in (\mathcal{X})^{N+2}$ defined as $\zeta = (x, \xi_0, \xi_1, \dots, \xi_{N-1}, x')$ satisfies

$$\mathcal{A}_N(\xi) \geq \mathcal{A}_{N+2}(\zeta) - 2\varepsilon \,. \tag{D.5.7}$$

and, by definition of $\widetilde{B}_{i,j}$, we thus obtain,

$$\mathcal{A}_N(\xi) \geq \mathcal{A}_{N+2}(\zeta) - 2\varepsilon \geq \widetilde{B}_{i,j} - 2\varepsilon \,. \tag{D.5.8}$$

Fix $x \in \mathcal{V}_i$. Let us now bound the probability

$$\mathbb{Q}_{\mathcal{V}}(x, \mathcal{V}_j) = \mathbb{P}_x\big(x^\eta_{\sigma_{\mathcal{V}}} \in \mathcal{V}_i\big) \,. \tag{D.5.9}$$

We have, for any $N \geq 0$,

$$\mathbb{P}_x\big(x^\eta_{\sigma_{\mathcal{V}}} \in \mathcal{V}_i\big) \leq \mathbb{P}_x\big(x^\eta_{\sigma_{\mathcal{V}}} \in \mathcal{V}_i,\, \sigma_{\mathcal{V}} < N\big) + \mathbb{P}_x(\sigma_{\mathcal{V}} \geq N) \,. \tag{D.5.10}$$

We first bound the second probability using Lemma D.21 applied to $\mathcal{U} \leftarrow \mathcal{V}_i$. Take N such that $\alpha_0(a - N) + \eta_0 b \leq -\widetilde{B}_{i,j}$. Then, by Markov's inequality and Lemma D.21, it holds that for all $\eta \leq \eta_0$

$$\begin{aligned}
\mathbb{P}_x(\sigma_{\mathcal{V}} \geq N) &\leq \mathbb{P}_x\left(\exp\left(\frac{\alpha_0 \sigma_{\mathcal{V}}}{\eta}\right) \geq \exp\left(\frac{\alpha_0 N}{\eta}\right)\right) \\
&\leq \exp\left(\frac{\alpha_0(a - N)}{\eta} + b\right) \\
&\leq \exp\left(\frac{-\widetilde{B}_{i,j}}{\eta}\right) .
\end{aligned} \tag{D.5.11}$$

We now bound the term $\mathbb{P}_x\big(x^\eta_{\sigma_{\mathcal{V}}} \in \mathcal{V}_j,\, \sigma_{\mathcal{V}} < N\big)$ for this choice of N.

For this, we show that $x^\eta_{\sigma_{\mathcal{V}}} \in \mathcal{V}_j$ with $\sigma_{\mathcal{V}} < N$ implies that

$$\operatorname{dist}_N\Big(x^\eta, \Gamma_N^{\{x\}}\Big(\widetilde{B}_{i,j} - 3\varepsilon\Big)\Big) > \frac{\delta'}{2} \,. \tag{D.5.12}$$

Indeed, on the event $x^\eta_{\sigma_{\mathcal{V}}} \in \mathcal{V}_j$ with $\sigma_{\mathcal{V}} < N$, there is some $N' \leq N$ such that $\sigma_{\mathcal{V}} = N' - 1$. If $\operatorname{dist}_N\Big(x^\eta, \Gamma_N^{\{x\}}\Big(\widetilde{B}_{i,j} - 3\varepsilon\Big)\Big) > \frac{\delta'}{2}$ did not hold, this would mean that there exists $\xi \in (\mathcal{X})^{N'}$ such that $\operatorname{dist}_{N'}(x^\eta, \xi) < \delta'$, $\xi_0 = x$ and, $\mathcal{A}_{N'}(\xi) \leq \widetilde{B}_{i,j} - 3\varepsilon$. In particular, ξ would also satisfy $\xi_{N'-1} \in \mathcal{U}_{\delta'}(\mathcal{V}_j)$, $\xi_n \in \mathcal{U}_{\delta'}(\mathcal{X} \setminus \mathcal{V})$ for all $n = 1, \dots, N' - 2$. This would be in direct contradiction of Eq. (D.5.8).

Therefore, we have that

$$\mathbb{P}_x\big(x^\eta_{\sigma_{\mathcal{V}}} \in \mathcal{V}_j,\, \sigma_{\mathcal{V}} < N\big) \leq \mathbb{P}_x\left(\operatorname{dist}_N\Big(x^\eta, \Gamma_N^{\{x\}}\Big(\widetilde{B}_{i,j} - 3\varepsilon\Big)\Big) > \frac{\delta'}{2}\right)$$

$$\leq \exp\left(-\frac{\widetilde{B}_{i,j} - 4\varepsilon}{\eta}\right), \tag{D.5.13}$$

by Corollary C.2.

Combining this bound with Eq. (D.5.11) yields

$$\mathbb{P}_x(x^\eta_{\sigma_\mathcal{V}} \in \mathcal{V}_j) \leq \exp\left(-\frac{\widetilde{B}_{i,j} - 4\varepsilon}{\eta}\right) + \exp\left(\frac{-\widetilde{B}_{i,j}}{\eta}\right), \tag{D.5.14}$$

which concludes the proof. ∎

Lemma D.23. *For any $\varepsilon > 0$,for any neighborhoods $\mathcal{V}_i$ of $\mathcal{K}_i$, $i = 1, \dots, K$ small enough, there exists $N \geq 0$, $\eta_0 > 0$ such that for all i, j, $x \in \mathcal{V}_i$, $0 < \eta < \eta_0$,*

$$\mathbb{Q}^N_\mathcal{V}(x, \mathcal{V}_j) \geq \exp\left(-\frac{\widetilde{B}_{i,j}}{\eta} - \frac{\varepsilon}{\eta}\right). \tag{D.5.15}$$

Proof. For any i, j, there exists $N_{i,j} \geq 1$, $\xi^{i,j} \in (\mathcal{X})^{N_{i,j}}$ such that $\xi^{i,j}_0 \in \mathcal{K}_i$, $\xi^{i,j}_{N_{i,j}-1} \in \mathcal{K}_j$, $\xi^{i,j}_n \notin \bigcup_{l \neq i,j} \mathcal{K}_l$ for all $n = 1, \dots, N_{i,j} - 2$ and $\mathcal{A}_{N_{i,j}}(\xi^{i,j}) \leq \widetilde{B}_{i,j} + \varepsilon$. Define $\delta_{i,j} \coloneqq \min\left\{d(\xi^{i,j}_n, \bigcup_{l \neq i,j} \mathcal{K}_l) : n = 1, \dots, N_{i,j} - 2\right\}$ and $\delta \coloneqq \min_{i,j \in I} \delta_{i,j}$. By construction, it holds that $\delta > 0$.

Require that $\mathcal{V}_i$ be contained in $\mathcal{W}_i \cap \mathcal{U}_{\delta/2}(\mathcal{K}_i)$ for all $i = 1, \dots, K$. Now, given such $\mathcal{V}_i$ neighborhoods of $\mathcal{K}_i$, $i = 1, \dots, K$, there exists $0 < \delta' \leq \delta/2$ such that $\mathcal{U}'_\delta(\mathcal{K}_i)$ is contained in $\mathcal{V}_i$ for all $i = 1, \dots, K$.

Apply Lemma D.9 to $\mathcal{K}_i$, $i = 1, \dots, K$ with $\varepsilon \leftarrow \min(\varepsilon, \delta'/2)$ and denote by N_i the bound on the length of paths obtained. Define

$$N \coloneqq \max_{i \in I} N_i + 1\,. \tag{D.5.16}$$

Fix $i, j \in I$ and $x \in \mathcal{V}_i$. Since $\mathcal{V}_i \subset \mathcal{W}_i$, there exists $z \in \mathcal{K}_i$ such that $\rho(x, z) < \varepsilon$. Moreover, note that $\rho(z, z) = 0$ since z is a critical point of f.

By Lemma D.9, there exists $n \leq N$, $\xi \in (\mathcal{X})^n$ such that $\xi_0 = z$, $\xi_{N-1} = \xi^{i,j}_0$, $\xi_k \in \mathcal{U}_{\delta'/2}(\mathcal{K}_i)$ for all $k = 1, \dots, n-2$ and $\mathcal{A}_n(\xi) < \varepsilon$.

Considering the concatenation

$$\zeta \coloneqq \left(x, \underbrace{z, z, \dots, z}_{N-n \text{ times}}, \xi_0, \xi_1, \dots, \xi_{n-2}, \xi_{n-1}, \xi^{i,j}_1, \dots, \xi^{i,j}_{N_{i,j}-1}\right) \tag{D.5.17}$$

which is a path of length $N + N_{i,j}$ made of $x \in \mathcal{V}_i$, then exactly N points in $\mathcal{U}_{\delta'/2}(\mathcal{K}_i)$ then $N_{i,j} - 2$ in $\mathcal{X} \setminus \mathcal{U}_{\delta/2}(\mathcal{V})$ and $\xi^{i,j}_{N_{i,j}-1} \in \mathcal{K}_j$. Moreover, by construction, $\mathcal{A}_{N+N_{i,j}}(\zeta) \leq \widetilde{B}_{i,j} + 3\varepsilon$. Therefore, if

$$\operatorname{dist}_{N+N_{i,j}}(x^\eta, \zeta) < \delta'/2\,, \tag{D.5.18}$$

with $x^\eta_0 = x$, then $x^\eta_1, \dots, x^\eta_N$ are in $\mathcal{U}_{\delta'}(\mathcal{K}_i) \subset \mathcal{V}_i$ and, since $\delta' \leq \delta/2$, $x^\eta_{N+1}, \dots, x^\eta_{N+N_{i,j}-2}$ are not in $\mathcal{U}_{\delta/4}(\mathcal{V})$, and therefore not in $\mathcal{V}$. Moreover, $x^\eta_{N+N_{i,j}-1}$ would be in $\mathcal{U}_{\delta'/2}(\mathcal{K}_j) \subset \mathcal{V}_j$.

Thus, all the paths x^η satisfying (D.5.18) with $x^\eta_0 = x$ are started at x and their N-th point that fall into $\mathcal{V}$ belongs to $\mathcal{V}_j$.

Therefore, using the definition of $\mathbb{Q}^N_\mathcal{V}$, we have that

$$\begin{aligned}\mathbb{Q}^N_\mathcal{V}(x, \mathcal{V}_j) &\geq \mathbb{P}_x\big(\operatorname{dist}_{N+N_{i,j}}(x^\eta, \zeta) < \delta'/2\big) \\ &\geq \exp\left(-\frac{\widetilde{B}_{i,j} + 4\varepsilon}{\eta}\right),\end{aligned} \tag{D.5.19}$$

by Corollary C.2. ∎

Lemma D.24. *For any $i, j \in I$,*

$$B_{i,j} = \min\left\{\sum_{l=0}^{n-2} \widetilde{B}_{i_l,i_{l+1}} : i_0 = i,\, i_{n-1} = j,\, i_l \in I \text{ for } l = 1,\dots,n-2\,, n \geq 1\right\}$$
$$= \min\left\{\sum_{l=0}^{K-2} \widetilde{B}_{i_l,i_{l+1}} : i_0 = i,\, i_{K-1} = j,\, i_l \in I \text{ for } l = 1,\dots,K-2\right\}. \tag{D.5.20}$$

Proof. It suffices to show that

$$B_{i,j} = \min\left\{\sum_{l=0}^{n-2} \widetilde{B}_{i_l,i_{l+1}} : i_0 = i,\, i_{n-1} = j,\, i_l \in I \text{ for } l = 1,\dots,n-2\,, n \geq 1\right\}. \tag{D.5.21}$$

The statement of the lemma then follows from the fact that $\widetilde{B}_{l,l} = 0$ for all $l \in I$ and that shortest paths on graphs can be chosen not to visit the same node twice.

For the inequality $(\geq)$, notice that any path between $\mathcal{K}_i$ and $\mathcal{K}_j$ can be decomposed into a concatenation of paths between $\mathcal{K}_i$ and $\mathcal{K}_{i_1}$, $\mathcal{K}_{i_1}$ and $\mathcal{K}_{i_2}$, …, $\mathcal{K}_{i_{n-1}}$ and $\mathcal{K}_j$ for some $i_1, \dots, i_{n-1} \in I$ that do not enter any other equivalence class in between. Therefore, the inequality $(\geq)$ follows from the definition of $B_{i,j}$ and $\widetilde{B}_{i_l,i_{l+1}}$.

We now focus on $(\leq)$.

Fix ε. Take $n \geq 1$, $i_0 = i$, $i_{n-1} = j$, $i_l \in I$ for $l = 1, \dots, n-2$. There are paths $\xi^0, \dots, \xi^{n-2}$ of lengths $N_0, \dots, N_{n-2}$ such that $\xi^l_0 \in \mathcal{K}_{i_l}$, $\xi^l_{N_l-1} \in \mathcal{K}_{i_{l+1}}$, $\mathcal{A}_{N_l}(\xi^l) \leq \widetilde{B}_{i_l,i_{l+1}} + \varepsilon/n$ for $l = 0, \dots, n-2$.

By Lemma D.9, for all $l = 0, \dots, n-2$, $\xi^l_{N_l-1}$ and ξ^{l+1}_0 can be connected by path of cost at most ε/n. Therefore concatenating all these paths yield ζ of length N with $\zeta_0 \in \mathcal{K}_i$, $\zeta_{N-1} \in \mathcal{K}_j$ and $\mathcal{A}_N(\zeta) \leq \sum_{l=0}^{n-2} \widetilde{B}_{i_l,i_{l+1}} + 2\varepsilon$. Since $\mathcal{A}_N(\zeta) \geq B_{i,j}$, we obtain the desired result. ∎

Notation 1. We will write, for non-decreasing $f\colon \mathbb{R} \to \mathbb{R}$,

$$a \asymp f(b \pm c) \iff f(b-c) \leq a \leq f(b+c)\,. \tag{D.5.22}$$

Proposition D.2. *For any $\varepsilon > 0$ and any small enough neighborhoods $\mathcal{V}_i$ of $\mathcal{K}_i$, $i = 1, \dots, K$, there exists $N \geq 0$, $\eta_0 > 0$ such that for all i, j, $x \in \mathcal{V}_i$, $0 < \eta < \eta_0$,*

$$\mathbb{Q}^N_{\mathcal{V}}(x, \mathcal{V}_j) \asymp \exp\left(-\frac{B_{i,j}}{\eta} \pm \frac{\varepsilon}{\eta}\right). \tag{D.5.23}$$

Proof. Let us first start with $(\geq)$.

Let N satisfy the conditions of Lemma D.23 and define $N' \coloneqq (K-1)N$. For any $i, j \in I$, by Lemma D.24, there exist $i_0 = i$, $i_{K-1} = j$, $i_l \in I$ for $l = 1, \dots, K-2$ such that

$$B_{i,j} = \sum_{l=0}^{K-2} \widetilde{B}_{i_l,i_{l+1}}\,. \tag{D.5.24}$$

Therefore, we have that the probability of reaching $\mathcal{V}_j$ from $x \in \mathcal{V}_i$ in $N' = (K-1)N$ steps is greater than the probability of reaching sequentially the $\mathcal{V}_{i_{l+1}}$ from any point $x \in \mathcal{V}_{i_l}$, i.e.,

$$\inf_{x \in \mathcal{V}_i} \mathbb{Q}^{N'}_{\mathcal{V}}(x, \mathcal{V}_j) \geq \prod_{l=0}^{K-2} \inf_{x \in \mathcal{V}_{i_l}} \mathbb{Q}^{N}_{\mathcal{V}}(x, \mathcal{V}_{i_{l+1}})$$

$$\geq \prod_{l=0}^{K-2} \exp\left(-\frac{\widetilde{B}_{i_l,i_{l+1}}}{\eta} - \frac{\varepsilon}{\eta}\right) = \exp\left(-\frac{B_{i,j}}{\eta} - \frac{(K-1)\varepsilon}{\eta}\right) \tag{D.5.25}$$

where we used Lemma D.23 to get the second inequality.

Now for the reverse inequality ($\leq$). Take $i, j \in I$ and denote $i_0 = i$, $i_{N'-1} = j$. By Lemma D.22, we have that

$$\begin{aligned}
\sup_{x\in\mathcal{V}_i} \mathbb{Q}_{\mathcal{V}}^{N'}(x, \mathcal{V}_j) &\leq \sum_{i_1,\dots,i_{K-2}\in I} \prod_{l=0}^{K-2} \sup_{x\in\mathcal{V}_{i_l}} \mathbb{Q}_{\mathcal{V}}^{N}(x, \mathcal{V}_{i_{l+1}}) \\
&\leq \sum_{i_1,\dots,i_{K-2}\in I} \prod_{l=0}^{K-2} \exp\left(-\frac{\widetilde{B}_{i_l,i_{l+1}}}{\eta} + \frac{\varepsilon}{\eta}\right) \\
&\leq \sum_{i_1,\dots,i_{K-2}\in I} \exp\left(-\frac{\sum_{l=0}^{N'} \widetilde{B}_{i_l,i_{l+1}}}{\eta} + \frac{(K-1)\varepsilon}{\eta}\right).
\end{aligned} \tag{D.5.26}$$

Using Lemma D.24, we obtain that

$$\begin{aligned}
\sup_{x\in\mathcal{V}_i} \mathbb{Q}_{\mathcal{V}}^{N'}(x, \mathcal{V}_j) &\leq \sum_{i_1,\dots,i_{K-2}\in I} \exp\left(-\frac{B_{i,j}}{\eta} + \frac{(K-1)\varepsilon}{\eta}\right) \\
&= K^{K-2} \exp\left(-\frac{B_{i,j}}{\eta} + \frac{(K-1)\varepsilon}{\eta}\right)
\end{aligned} \tag{D.5.27}$$

which concludes the proof. ■

Let us now give our first estimates on the invariant measure.

Definition 8 ([20, Chap. 6,§4])**.** Define, for every $i \in I$,

$$E_i = E(\mathcal{K}_i) \coloneqq \min_{T\in\mathcal{T}_i} \sum_{(j\to l)\in T} B_{j,l}\,, \tag{D.5.28}$$

where $\mathcal{T}_i$ denotes the set of trees rooted at i in the complete graph on I.

Proposition D.3. *For any $\mu_{\mathcal{V}}$ invariant probability measure for $\mathbb{Q}_{\mathcal{V}}$, the induced chain on $\mathcal{V}$, in the setting of Proposition D.2, for any $i \in I$,*

$$\mu_{\mathcal{V}}(\mathcal{V}_i) \asymp \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j\in I} E(\mathcal{K}_j)}{\eta} \pm \frac{\varepsilon}{\eta}\right). \tag{D.5.29}$$

Proof. If $\mu_{\mathcal{V}}$ is an invariant measure of $\mathbb{Q}_{\mathcal{V}}$, then it is an invariant measure of $\mathbb{Q}_{\mathcal{V}}^N$ for any N given by Proposition D.2.

Freidlin & Wentzell [20, Chap .6, Lem. 3.1-3.2] combined with Proposition D.2 then give

$$\begin{aligned}
&\exp(-(2\operatorname{card} I + 2)\varepsilon/\eta) \frac{\exp(-E(\mathcal{K}_i)/\eta)}{\sum_{j\in I}\exp(-E(\mathcal{K}_j)/\eta)} \\
&\quad\leq \mu_{\mathcal{V}}(\mathcal{V}_i) \\
&\quad\leq \exp((2\operatorname{card} I + 2)\varepsilon/\eta) \frac{\exp(-E(\mathcal{K}_i)/\eta)}{\sum_{j\in I}\exp(-E(\mathcal{K}_j)/\eta)}\,.
\end{aligned} \tag{D.5.30}$$

For η small enough, it holds that

$$\sum_{j\in I} \exp(-E(\mathcal{K}_j)/\eta) \asymp \exp\left(-\min_{j\in I} E(\mathcal{K}_j)/\eta \pm \varepsilon/\eta\right), \tag{D.5.31}$$

which concludes the proof. ■

We now state a result that links invariant measures of $(x_n^\eta)_n$ and $\mathbb{Q}_\mathcal{V}$. It is a consequence of Douc et al. [14, Thm. 3.6.5].

Lemma D.25. *There is $\eta_0 > 0$ such that, for $0 < \eta \leq \eta_0$, if $(x_n^\eta)_n$ has an invariant probability measure μ_∞, then, for any $\mathcal{V}$ measurable neighborhood of* crit f*, we have that $\mu_\infty(\mathcal{V}) > 0$ and $\mu_\mathcal{V}$, the restriction of $\mu_\infty/\mu_\infty(\mathcal{V})$ to $\mathcal{V}$ is an invariant measure for the induced chain on $\mathcal{V}$ and, for any measurable set $E \subset \mathcal{X}$,*

$$\frac{\mu_\infty(E)}{\mu_\infty(\mathcal{V})} = \int_\mathcal{V} d\mu_\mathcal{V}(x)\, \mathbb{E}_x\left[\sum_{n=0}^{\sigma_\mathcal{V}-1} \mathbb{1}\{x_n^\eta \in E\}\right]. \tag{D.5.32}$$

Proof. We invoke the first item of Lemma D.19: there exists $\mathcal{D} \subset \mathcal{X}$ a compact set, $\eta_0 > 0$, such that for any $\mathcal{D}' \subset \mathcal{X}$ compact set such that $\mathcal{D} \subset \mathcal{D}'$, there exists $\alpha_0 > 0$ such that,

$$\forall \eta \leq \eta_0,\, x \in \mathcal{D}', \quad \mathbb{E}_x\left[e^{\frac{\alpha_0 \sigma_\mathcal{D}}{\eta}}\right] < +\infty\,. \tag{D.5.33}$$

Without loss of generality, at the potential expense of expanding $\mathcal{D}$, assume that $\mathcal{V} \subset \mathcal{D}$.

By applying Lemma D.21 to $\mathcal{U} \leftarrow \mathcal{V}$ and $\mathcal{D} \leftarrow \mathcal{D}$, we get that there is some $\eta_0, \alpha_0, > 0$ such that, for any $\eta \leq \eta_0$, $x \in \mathcal{D}$,

$$\mathbb{E}_x\left[e^{\frac{\alpha \sigma_\mathcal{V}}{\eta}}\right] < +\infty\,. \tag{D.5.34}$$

In particular, we have that $\mathbb{P}_x(\sigma_\mathcal{V} < \infty) = 1$ for any $x \in \mathcal{D}$, and *a fortiori* for any $x \in \mathcal{V}$.

Let us now show that, for any $x \in \mathcal{X}$, $\mathbb{P}_x(\sigma_\mathcal{V} < \infty) = 1$. Fix $x \in \mathcal{X}$. By choosing $\mathcal{D}'$ large enough to contain both $\mathcal{D}$ and x, Eq. (D.5.33) implies that, with $x_0 = x$, $\sigma_\mathcal{D} < \infty$ almost surely. Therefore, by the strong Markov property, it holds that,

$$\mathbb{P}_x(\sigma_\mathcal{V} < \infty) \geq \inf_{z \in \mathcal{D}} \mathbb{P}_z(\sigma_\mathcal{V} < \infty) = 1\,. \tag{D.5.35}$$

Therefore the assumptions of Douc et al. [14, Thm. 3.6.5] are satisfied as well as its first item which yields the result. ∎

We reach the next proposition, which is the first part of our main result. It is an adaptation of Freidlin & Wentzell [20, Thm. 4.1] to the discrete time setting.

Proposition D.4. *For any $\varepsilon > 0$, for any $\mathcal{V}_1, \dots, \mathcal{V}_K$ measurable neighborhoods of $\mathcal{K}_1, \dots, \mathcal{K}_K$ small enough, there exists $\eta_0 > 0$ such that for any $0 < \eta < \eta_0$, μ_∞ invariant probability measure for $(x_n^\eta)_n$, for any $i \in I$,*

$$\mu_\infty(\mathcal{V}_i) \asymp \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j \in I} E(\mathcal{K}_j)}{\eta} \pm \frac{\varepsilon}{\eta}\right). \tag{D.5.36}$$

Proof. Let us first provide estimates for the unnormalized measure defined by, for any measurable set E,

$$\widetilde{\mu}(E) \coloneqq \int_E d\mu_\mathcal{V}(x)\, \mathbb{E}_x\left[\sum_{n=0}^{\sigma_\mathcal{V}-1} \mathbb{1}\{x_n^\eta \in E\}\right]. \tag{D.5.37}$$

By definition of $\sigma_\mathcal{V}$, in the sequence of points $x_0^\eta, \dots, x_{\sigma_\mathcal{V}-1}^\eta$, only x_0^η can be in $\mathcal{V}$. Therefore, for any $i, j \in I$, $x \in \mathcal{V}_j$,

$$\mathbb{E}_x\left[\sum_{n=0}^{\sigma_\mathcal{V}-1} \mathbb{1}\{x_n^\eta \in \mathcal{V}_i\}\right] = \begin{cases} 1 & \text{if } i = j\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{D.5.38}$$

Thus, we have that,

$$\widetilde{\mu}(\mathcal{V}_i) = \mu_\mathcal{V}(\mathcal{V}_i)$$

$$\asymp \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j\in I} E(\mathcal{K}_j)}{\eta} \pm \frac{\varepsilon}{\eta}\right), \tag{D.5.39}$$

by Proposition D.3. It now remains to estimate the normalization constant $\widetilde{\mu}(\mathcal{X})$. On the one hand, we have that

$$\begin{aligned}\widetilde{\mu}(\mathcal{X}) &\geq \max_{i\in I} \widetilde{\mu}(\mathcal{V}_i) \\ &\geq \max_{i\in I} \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j\in I} E(\mathcal{K}_j)}{\eta} - \frac{\varepsilon}{\eta}\right) = \exp\left(-\frac{\varepsilon}{\eta}\right).\end{aligned} \tag{D.5.40}$$

On the other hand, by Lemma D.21 applied with $\mathcal{U} \leftarrow \mathcal{V}$ and $\mathcal{D} \leftarrow \operatorname{cl}\mathcal{V}$ (choosing $\mathcal{V}$ small enough so that is bounded), the quantity

$$c \coloneqq \sup\{\mathbb{E}_x[\sigma_{\mathcal{V}}] : x \in \mathcal{V},\, 0 < \eta \leq \eta_0\} \tag{D.5.41}$$

is finite. Therefore, we have that

$$\begin{aligned}\widetilde{\mu}(\mathcal{X}) &= \int_{\mathcal{X}} d\mu_{\mathcal{V}}(x)\, \mathbb{E}_x[\sigma_{\mathcal{V}}] \\ &\leq c \int_{\mathcal{X}} d\mu_{\mathcal{V}}(x) = c\,,\end{aligned} \tag{D.5.42}$$

which, along with choosing η_0 small enough so that $c \leq \exp(\varepsilon/\eta)$, concludes the proof. ■

We will need the following lemma to prove the second part of our main result.

Lemma D.26. *For any $\varepsilon > 0$, for any $\mathcal{V}_1, \dots, \mathcal{V}_K$ measurable neighborhoods of $\mathcal{K}_1, \dots, \mathcal{K}_K$ small enough, D measurable set, $\delta_D > 0$, there exists $\eta_0 > 0$ such that, for any $0 < \eta < \eta_0$, for any $i \in I$, $x \in \mathcal{V}_i$,*

$$\mathbb{P}_x\left(\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}\right) \leq \exp\left(-\frac{B(\mathcal{K}_i, D) - \varepsilon}{\eta}\right), \tag{D.5.43}$$

where

$$D_{-\delta_D} \coloneqq \{z \in D : d(z, \mathcal{X} \setminus D) \geq \delta_D\}. \tag{D.5.44}$$

Proof. The requirement on the $\mathcal{V}_i$ are the same in the proof of Lemma D.22 but we restate them here for completeness.

Assume that, without loss of generality, ε is small enough so that Lemma D.8 with $r \leftarrow \varepsilon$ can be applied to every $\mathcal{K}_i$, $i = 1, \dots, K$. Denote by $\mathcal{W}_i$, $i = 1, \dots, K$ the corresponding neighborhoods of $\mathcal{K}_i$. Since these $\mathcal{W}_i$'s are neighborhoods of the $\mathcal{K}_i$'s, there exists $\delta > 0$ such that $\mathcal{U}_{2\delta}(\mathcal{K}_i) \subset \mathcal{W}_i$ for all $i = 1, \dots, K$. Require then that $\mathcal{V}_i$ be contained in $\mathcal{U}_\delta(\mathcal{K}_i)$ so that $\mathcal{U}_\delta(\operatorname{cl}\mathcal{V}_i) \subset \mathcal{W}_i$ for all $i = 1, \dots, K$. Moreover, assume that $\delta > 0$ is small enough so that the neighborhoods $\mathcal{U}_\delta(\mathcal{K}_i)$, $i = 1, \dots, K$ are pairwise disjoint and that $\delta \leq \delta_D$. Define $0 < \delta' \leq \delta$ such that $\mathcal{U}'_\delta(\mathcal{K}_i)$ is contained in $\mathcal{V}_i$ for all $i = 1, \dots, K$.

Fix $i \in I$ and consider $\xi \in (\mathcal{X})^N$ such that $\xi_0 \in \mathcal{U}'_\delta(\operatorname{cl}\mathcal{V}_i)$, $\xi_{N-1} \in \mathcal{U}'_\delta(D_{-\delta_D})$. By construction, $\xi_0 \in \mathcal{U}_\delta(\mathcal{V}_i) \subset \mathcal{W}_i$. Therefore, there exists $x \in \mathcal{K}_i$ such that $\rho(x, \xi_0) < \varepsilon$. Moreover, since $\delta' \leq \delta \leq \delta_D$, $\mathcal{U}'_\delta(D_{-\delta_D}) \subset D$ so that $\xi_{N-1} \in D$.

Define $\zeta \coloneqq (x, \xi_0, \dots, \xi_{N-1})$ which is a path from $\mathcal{K}_i$ to D so that

$$\mathcal{A}_N(\xi) \geq \mathcal{A}_{N+1}(\zeta) - \varepsilon \geq B(\mathcal{K}_i, D) - \varepsilon\,. \tag{D.5.45}$$

We now follow the same outline as for the proof of Lemma D.22. Fix $x \in \mathcal{V}_i$. For any $N \geq 0$, we have that

$$\mathbb{P}_x\left(\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}\right) \leq \mathbb{P}_x\left(\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}\,, \sigma_{\mathcal{V}} < N\right) + \mathbb{P}_x(\sigma_{\mathcal{V}} \geq N)\,. \tag{D.5.46}$$

For some N large enough, Lemma D.21 yields

$$\mathbb{P}_x(\sigma_{\mathcal{V}} \geq N) \leq \exp\left(-\frac{B(\mathcal{K}_i, D) - \varepsilon}{\eta}\right). \tag{D.5.47}$$

We now bound $\mathbb{P}_x\left(\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}, \sigma_{\mathcal{V}} < N\right)$ for this choice of N. For this, it suffices to note that $\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}, \sigma_{\mathcal{V}} < N$ implies that $\operatorname{dist}_N\left(x^\eta, \Gamma_N^{\{x\}}(\xi - 2\varepsilon)\right) > \frac{\delta'}{2}$. Then, applying Corollary C.2 yields

$$\mathbb{P}_x\left(\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}, \sigma_{\mathcal{V}} < N\right) \leq \exp\left(-\frac{B(\mathcal{K}_i, D) - 3\varepsilon}{\eta}\right), \tag{D.5.48}$$

which concludes the proof. ∎

We can use this lemma to upper-bound the $\mu_\infty(D)$.

Lemma D.27. *For any $\varepsilon > 0$, any bounded measurable set D, there exists $\eta_0 > 0$ such that for any $0 < \eta < \eta_0$, μ_∞ invariant probability measure for $(x_n^\eta)_n$, for any $i \in I$,*

$$\mu_\infty(D_{-\delta_D}) \leq \exp\left(-\frac{\min_{i\in I}\{E(\mathcal{K}_i) + B(\mathcal{K}_i, D)\} - \min_{i\in I} E(\mathcal{K}_i)}{\eta} + \frac{\varepsilon}{\eta}\right), \tag{D.5.49}$$

where

$$D_{-\delta_D} \coloneqq \{z \in \mathcal{X} : d(z, \mathcal{X} \setminus D) \geq \delta_D\}. \tag{D.5.50}$$

Proof. Using $\mathcal{V}_1, \dots, \mathcal{V}_K$ measurable neighborhoods of $\mathcal{K}_1, \dots, \mathcal{K}_K$ small enough given by Lemma D.26, we provide an estimate for the weight of D for the unnormalized measure $\widetilde{\mu}$ defined in the proof of Proposition D.4, i.e., for

$$\widetilde{\mu}(D_{-\delta_D}) \coloneqq \int_{\mathcal{V}} d\mu_{\mathcal{V}}(x)\, \mathbb{E}_x\left[\sum_{n=0}^{\sigma_{\mathcal{V}}-1} \mathbb{1}\{x_n^\eta \in D_{-\delta_D}\}\right], \tag{D.5.51}$$

and the result will follow from the estimate on the normalization constant $\widetilde{\mu}(\mathcal{X})$ obtained in the proof of Proposition D.4.

By Lemma D.21 applied with $\mathcal{U} \leftarrow \mathcal{V}$ and $\mathcal{D} \leftarrow \operatorname{cl}\mathcal{V} \cup D$ (choosing $\mathcal{V}$ small enough so that is bounded), the quantity

$$c \coloneqq \sup\{\mathbb{E}_x[\sigma_{\mathcal{V}}] : x \in \mathcal{V} \cup D,\, 0 < \eta \leq \eta_0\} \tag{D.5.52}$$

is finite.

Fix $i \in I$ and $x \in \mathcal{V}_i$. If $\sigma_{D_{-\delta_D}} \geq \sigma_{\mathcal{V}}$, then $\sum_{n=0}^{\sigma_{\mathcal{V}}-1} \mathbb{1}\{x_n^\eta \in D_{-\delta_D}\}$ would be 0 so, we have that

$$\begin{aligned}
\mathbb{E}_x\left[\sum_{n=0}^{\sigma_{\mathcal{V}}-1} \mathbb{1}\{x_n^\eta \in D_{-\delta_D}\}\right] &= \mathbb{E}_x\left[\mathbb{1}\{\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}\} \sum_{n=0}^{\sigma_{\mathcal{V}}-1} \mathbb{1}\{x_n^\eta \in D_{-\delta_D}\}\right] \\
&\leq \mathbb{E}_x\left[\mathbb{1}\{\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}\}\sigma_{\mathcal{V}}\right] \\
&= \mathbb{E}_x\left[\mathbb{1}\{\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}\}\left(\sigma_{D_{-\delta_D}} + \mathbb{E}_{x^\eta_{\sigma_{D_{-\delta_D}}}}[\sigma_{\mathcal{V}}]\right)\right],
\end{aligned} \tag{D.5.53}$$

where we used the strong Markov property. Bounding $\sigma_{D_{-\delta_D}}$ by $\sigma_{\mathcal{V}}$ on the event $\{\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}\}$, we obtain that

$$\begin{aligned}
\mathbb{E}_x\left[\sum_{n=0}^{\sigma_{\mathcal{V}}-1} \mathbb{1}\{x_n^\eta \in D_{-\delta_D}\}\right] &\leq \mathbb{E}_x\left[\mathbb{1}\{\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}\}\left(\sigma_{\mathcal{V}} + \mathbb{E}_{x^\eta_{\sigma_{\mathcal{V}}}}[\sigma_{\mathcal{V}}]\right)\right] \\
&\leq 2c\,\mathbb{P}_x\left(\sigma_{D_{-\delta_D}} < \sigma_{\mathcal{V}}\right)
\end{aligned}$$

$$\leq 2c \exp\left(-\frac{B(\mathcal{K}_i, D) - \varepsilon}{\eta}\right), \tag{D.5.54}$$

where we invoked Lemma D.26 for the last inequality.

Combining this bound and Proposition D.3, we obtain that

$$\begin{aligned}\widetilde{\mu}(D_{-\delta_D}) &\leq 2c \sum_{i \in I} \mu_{\mathcal{V}}(\mathcal{V}_i) \exp\left(-\frac{B(\mathcal{K}_i, D) - \varepsilon}{\eta}\right) \\ &\leq 2c \operatorname{card} I \exp\left(-\frac{\min_{i \in I}\{E(\mathcal{K}_i) + B(\mathcal{K}_i, D)\} - \min_{i \in I} E(\mathcal{K}_i)}{\eta} + \frac{2\varepsilon}{\eta}\right), \end{aligned} \tag{D.5.55}$$

which concludes the proof. ■

D.6. **Convergence and stability.** Let us first begin by showing that, for every initial point, the flow of f converges to one of the $\mathcal{K}_i$.

Lemma D.28. *For any $x \in \mathcal{X}$, there is $i \in I$ such that*

$$\lim_{t \to +\infty} d(\Theta_t(x), \mathcal{K}_i) = 0. \tag{D.6.1}$$

Proof. Fix $x \in \mathcal{X}$. For any $t \geq 0$, $\Theta_t(x)$ belongs to $\{z \in \mathcal{X} : f(z) \leq f(x)\}$, which is compact by coercivity of f. Therefore, the set of accumulation points of $(\Theta_t(x))_{t \geq 0}$ is non-empty, connected and included in $\operatorname{crit} f$. Therefore, since the $\mathcal{K}_i$, for $i \in I$, are the connected components of $\operatorname{crit} f$, there is $i \in I$ such that the accumulation points of $(\Theta_t(x))_{t \geq 0}$ all belong to $\mathcal{K}_i$.

If $d(\Theta_t(x), \mathcal{K}_i)$ did not converge to 0, there would be a subsequence that would converge to some point out of any $\mathcal{K}_i$, which would be a contradiction. Therefore, $d(\Theta_t(x), \mathcal{K}_i)$ must converge to 0. ■

Let us restate the definition of minimizing component that we introduced in the main text.

Definition 9. For any $i \in I$, we say that $\mathcal{K}_i$ is a *minimizing component* if there exists $\mathcal{U}$ a neighborhood of $\mathcal{K}_i$ such that,

$$\operatorname*{arg\,min}_{x \in \mathcal{U}} f(x) = \mathcal{K}_i \,. \tag{D.6.2}$$

$\mathcal{K}_i$ is called minimizing otherwise.

We now state a standard definition for asymptotic stability.

Definition 10. A connected component of the critical points $\mathcal{K}_i$, for some $i \in I$, is said to be *asymptotically stable* if there exists $\mathcal{U}$ a neighborhood of $\mathcal{K}_i$ such that, for any $x \in \mathcal{U}$, $\Theta_t(x)$ converges to $\mathcal{K}_i$.

The notions of minimizing component and asymptotic stability are equivalent in our context.

Lemma D.29. *For any $i \in I$, $\mathcal{K}_i$ is a minimizing component if and only if it is asymptotically stable.*

Proof. We start with the direct implication. Assume that $\mathcal{K}_i$ is a minimizing component. Therefore, there exists $\delta > 0$ such that

$$\operatorname*{arg\,min}_{x \in \operatorname{cl} \mathcal{U}_\delta(\mathcal{K}_i)} f(x) = \mathcal{K}_i \,. \tag{D.6.3}$$

Moreover, assume that δ is small enough so that, $\operatorname{cl} \mathcal{U}_\delta(\mathcal{K}_i) \cap \operatorname{crit} f = \mathcal{K}_i$.

Then, for any $x \in \mathcal{U}_\delta(\mathcal{K}_i)$, $f(x) < f(\mathcal{K}_i)$ and, for any $x \in \mathcal{X} \setminus \mathrm{cl}\,\mathcal{U}_\delta(\mathcal{K}_i)$, $f(x) > f(\mathcal{K}_i)$. By continuity of f, compactness of $\mathrm{cl}\,\mathcal{U}_\delta(\mathcal{K}_i) \setminus \mathcal{U}_{\delta/2}(\mathcal{K}_i)$ and definition of minimizing component, we have that

$$\min_{x \in \mathrm{cl}\,\mathcal{U}_\delta(\mathcal{K}_i) \setminus \mathcal{U}_{\delta/2}(\mathcal{K}_i)} f(x) > f(\mathcal{K}_i)\,. \tag{D.6.4}$$

Define $\delta' \coloneqq \frac{1}{2} \min_{x \in \mathrm{cl}\,\mathcal{U}_\delta(\mathcal{K}_i) \setminus \mathcal{U}_{\delta/2}(\mathcal{K}_i)} \{f(x) - f(\mathcal{K}_i)\}$ and

$$\mathcal{V} \coloneqq \left\{x \in \mathcal{U}_{\delta/2}(\mathcal{K}_i) : f(x) < f(\mathcal{K}_i) + \delta'\right\}, \tag{D.6.5}$$

which is a neighborhood of $\mathcal{K}_i$ by continuity of f.

We now show that trajectories of the flow starting in $\mathcal{V}$ converge to $\mathcal{K}_i$. Take $x \in \mathcal{V}$. Since $f(\Theta_t(x))$ is non-increasing, then $(\Theta_t(x))_t$ remains in $\mathcal{V}$ by construction. By Lemma D.28, $(\Theta_t(x))_t$ converges to some component of the critical points, which must be $\mathcal{K}_i$ since $\mathcal{V}$ is disjoint from the other components.

We now show the converse implication. Since $\mathcal{K}_i$ is a connected component of $\mathrm{crit}\, f$, f is constant on $\mathcal{K}_i$. Denote by f_i^* this value. Take $\delta > 0$ small enough such that $\mathcal{U}_\delta(\mathcal{K}_i) \cap \mathrm{crit}\, f = \mathcal{K}_i$ and such that, for any $x \in \mathcal{U}_\delta(\mathcal{K}_i)$, $(\Theta_t(x))_t$ converges to $\mathcal{K}_i$. Take $x \in \mathcal{U}_\delta(\mathcal{K}_i) \setminus \mathcal{K}_i$. We show that $f(x) > f_i^*$. For any $t > 0$,

$$f(\Theta_t(x)) - f(\mathcal{K}_i) = -\int_0^t \|\nabla f(\Theta_s(x))\|^2 \, ds\,, \tag{D.6.6}$$

which must (strictly) negative since x is not a critical point of f. Since $f(\Theta_t(x))$ is non-increasing in t and lower-bounded — because $\inf_{\mathcal{X}} f > -\infty$ by coercivity of f —, it must converge as $t \to +\infty$ and its limit satisfies $\lim_{t\to+\infty} f(\Theta_t(x)) < f(x)$. Moreover, $(\Theta_t(x))_t$ converges to $\mathcal{K}_i$ so that all its accumulation points belong to $\mathcal{K}_i$ and have the same objective value f_i^*. Hence, we have that $\lim_{t\to+\infty} f(\Theta_t(x)) = f_i^*$ and $f_i^* < f(x)$. ∎

In the following, we will thus use the terms minimizing component and asymptotically stable interchangeably.

The next lemma shows that if $\mathcal{K}_i$ is not asymptotically stable, then it is unstable in the sense of Freidlin & Wentzell [20, Chap. 6,§4].

Lemma D.30. *If $\mathcal{K}_i$ is not asymptotically stable, then there exists $j \in I$ such that $B_{i,j} = 0$.*

Proof. If $\mathcal{K}_i$ is not asymptotically stable, then, for every $n \geq 1$, there exists $x_n \in \mathcal{U}_{1/n}(\mathcal{K}_i)$ such that $\Theta_t(x_n)$ does not converge to $\mathcal{K}_i$. By Lemma D.28, there exists $j_n \in I$ such that $(\Theta_t(x_n))_t$ converges to $\mathcal{K}_{j_n}$. Since I is finite, there exists $j \in I$ such that $j_n = j$ for infinitely many n. Replacing $(x_n)_{n\geq 1}$ by a subsequence, we can assume that $(\Theta_t(x_n))_t$ converges to $\mathcal{K}_j$ for all $n \geq 1$.

We now show that $B_{i,j} = 0$.

Fix $\varepsilon > 0$ and consider $\mathcal{W}_\varepsilon(\mathcal{K}_i), \mathcal{W}_\varepsilon(\mathcal{K}_j)$ which are neighborhoods of $\mathcal{K}_i, \mathcal{K}_j$ by Lemma D.8. Since $(x_n)_{n\geq 1}$ converges to $\mathcal{K}_i$, there exists n such that $x_n \in \mathcal{W}_\varepsilon(\mathcal{K}_i)$. In turn, since $\Theta_t(x_n)$ converges to $\mathcal{K}_j$, there exists $T > 0$, which we can choose integer, such that, $\Theta_T(x_n) \in \mathcal{W}_\varepsilon(\mathcal{K}_j)$. Therefore, there exists $x \in \mathcal{K}_i$ and $z \in \mathcal{K}_j$ such that $\rho(x, x_n) < \varepsilon$ and $\rho(\Theta_T(x_n), z) < \varepsilon$. Consider the discrete path $\xi \in \mathcal{X}^{T+3}$ defined by $\xi_0 = x$, $\xi_n = \Theta_{n-1}(x)$ for all $1 \leq n \leq T+1$ and $\xi_{T+2} = z$. By Lemma D.2, we have that

$$\mathcal{A}_{T+3}(\xi) = \rho(x, x_n) + \rho(\Theta_T(x), z) < 2\varepsilon\,, \tag{D.6.7}$$

and therefore, since $\xi_0 \in \mathcal{K}_i$ and $\xi_{T+2} \in \mathcal{K}_j$, $B_{i,j} < 2\varepsilon$. ∎

The following lemma is now a straightforward adaptation of Freidlin & Wentzell [20, Chap. 6, Lemma 4.2].

Lemma D.31. *If $\mathcal{K}_i$ is not asymptotically stable, then there exists $j \in I$ such that $B_{i,j} = 0$ and such that, for any $l \in I$, $B_{j,l} > 0$.*

Proof. For the sake of contradiction, assume that such a j does not exist. Then we build an infinite sequence $j_0 = i, j_1, \dots$ such that $B_{j_n, j_{n+1}} = 0$ for all $n \geq 0$. By definition of equivalence classes, any j cannot appear twice in this sequence. But since I is finite, this is a contradiction. ∎

We now reach our final result on unstable equivalence classes: they have negligible weight in the invariant measure.

Lemma D.32. *If $\mathcal{K}_i$ is not asymptotically stable, or, equivalently, non-minimizing, then there exists $j \in I$ such that $\mathcal{K}_j$ is asymptotically stable, $B_{i,j} = 0$ and,*

$$E_j < E_i \tag{D.6.8}$$

Proof. By Lemma D.31, there exists $j \in I$ such that $B_{i,j} = 0$ and such that, for any $l \in I$, $B_{j,l} > 0$. Lemma D.30 implies in particular that $\mathcal{K}_j$ must be asymptotically stable

It remains to show that $E_j < E_i$. Take $T \in \mathcal{T}_i$ such that

$$E_i = \sum_{(l \to k) \in T} B_{l,k} \,. \tag{D.6.9}$$

j has an outgoing edge in T and denote by j' the other end of that edge. Now, consider the tree $T' \in \mathcal{T}_j$ obtained from T by removing the outgoing edge from j to i' and adding an edge from i to j. Then, by definition of E_j,

$$E_j \leq \sum_{(l \to k) \in T'} B_{l,k} = \sum_{(l \to k) \in T} B_{l,k} - B_{j,j'} + B_{i,j} \,. \tag{D.6.10}$$

But, by definition of j, $B_{i,j} = 0$ and $B_{j,j'} > 0$. Therefore, $E_j < E_i$. ∎

We now show the second part of our main result: the invariant measure concentrates on the ground states, which are asymptotically stable by Lemma D.32.

For this, we need the following lemma.

Lemma D.33. *For any $i \in I$ such that $\mathcal{K}_i$ is minimizing, there exists $\delta > 0$ such that, for any $0 < \delta' \leq \delta$,*

$$B(\mathcal{K}_i, \mathcal{X} \setminus \mathcal{U}'_\delta(\mathcal{K}_i)) > 0 \,. \tag{D.6.11}$$

Proof. Since $\mathcal{K}_i$ is minimizing, there exists $\delta > 0$ such that, for any $x \in \mathcal{U}_\delta(\mathcal{K}_i)$, $f(x) > f_i$ where f_i is the value of f on $\mathcal{K}_i$. Take $\delta' \leq \delta$, $\mathcal{U} \coloneqq \mathcal{U}'_\delta(\mathcal{K}_i)$ and

$$\Delta \coloneqq \min\{U_\infty(x) - \alpha_\infty(f_i) : x \in \mathcal{X},\, d(x, \mathcal{K}_i) = \delta'/2\} \,. \tag{D.6.12}$$

Then, by the continuity of U_∞ and the fact that α_∞ is (strictly) increasing, we have that $\Delta > 0$. To conclude the proof of this lemma, we now show that $B(\mathcal{K}_i, \mathcal{X} \setminus \mathcal{U}) \geq \frac{\Delta}{2}$. Consider some $T > 0$ and $\gamma \in \mathcal{C}([0,T], \mathcal{X})$ such that $\gamma_0 \in \mathcal{K}_i$ and $\gamma_T \in \mathcal{X} \setminus \mathcal{W}$. By continuity of γ and $d(\cdot, \mathcal{K}_i)$, there exists $t \in [0,T]$ such that $d(\gamma_t, \mathcal{K}_i) = \delta'/2$. By the same computation as in Lemma D.6, we have that

$$\begin{aligned}
\Delta &\leq U_\infty(\gamma_t) - U_\infty(\gamma_0) \\
&= 2 \int_0^t \frac{\langle \dot{\gamma}_s, \nabla f(\gamma_s) \rangle}{\sigma_\infty^2(f(\gamma_s))} \\
&\leq 2\mathcal{S}_{[0,t]}(\gamma) \\
&\leq 2\mathcal{S}_{[0,T]}(\gamma) \,.
\end{aligned} \tag{D.6.13}$$

Since this is valid for any γ, we obtain that $B(\mathcal{K}_i, \mathcal{X} \setminus \mathcal{U}) \geq \frac{\Delta}{2}$. ∎

The next proposition shows that the invariant measure concentrates exponentially on states that are asymptotically stable (and contain the ground states).

Proposition D.5. *Consider $J \subset I$ such that, for all $i \in J$, $\mathcal{K}_i$ is minimizing and, such that, J contains* $\arg\min_{i\in I} E_i$. *Consider $\mathcal{V}_i$ small enough neighborhoods of $\mathcal{K}_i$ for $i \in J$. Then, there exists $c > 0$, $\eta_0 > 0$ such that, for any $\eta \leq \eta_0$, for any μ_∞ invariant measure of $(x_n^\eta)_{n\geq 0}$,*

$$\mu_\infty\left(\mathcal{X} \setminus \bigcup_{i\in J} \mathcal{V}_i\right) \leq e^{-\frac{c}{\eta}}\,. \tag{D.6.14}$$

Proof. Take $\delta > 0$ small enough so that, for any $i \in J$, $\mathcal{U}_\delta(\mathcal{K}_i) \subset \mathcal{V}_i$ and

$$d(\mathcal{K}_i, \mathcal{X} \setminus \mathcal{U}_\delta(\mathcal{K}_i)) > 0\,. \tag{D.6.15}$$

This is possible by Lemma D.33.

Define

$$D \coloneqq \mathcal{X} \setminus \bigcup_{i\in J} \mathcal{U}_{\delta/2}(\mathcal{K}_i)\,. \tag{D.6.16}$$

With the notations of Lemma D.27, we show that

$$\mathcal{X} \setminus \bigcup_{i\in J} \mathcal{V}_i \subset D_{-\delta/2}\,. \tag{D.6.17}$$

Indeed, take $x \in \mathcal{X} \setminus \bigcup_{i\in J} \mathcal{V}_i$. Then, for any $i \in J$, it holds that $d(x, \mathcal{K}_i) \geq \delta$ and so, we have

$$d(x, \mathcal{X} \setminus D) = d\left(x, \mathcal{X} \setminus \bigcup_{i\in J} \mathcal{U}_{\delta/2}(\mathcal{K}_i)\right) \geq \frac{\delta}{2}\,. \tag{D.6.18}$$

Hence $x \in D_{-\delta/2}$ and, it suffices to bound $\mu_\infty(D_{-\delta/2})$ to show the result.

Moreover, for any $i \in J$,

$$B(\mathcal{K}_i, D) \geq B(\mathcal{K}_i, \mathcal{X} \setminus \mathcal{U}_{\delta/2}(\mathcal{K}_i)) > 0\,, \tag{D.6.19}$$

so that the quantity

$$c \coloneqq \min\left(\min_{i\notin J} E_i - \min_{i\in I} E_i, \min_{i\in J} B(\mathcal{K}_i, D)\right) \tag{D.6.20}$$

is positive.

Apply Lemma D.27 with $\delta_D \leftarrow \delta/2$, $\varepsilon \leftarrow c/2$ to get that, for any $\eta \leq \eta_0$,

$$\mu_\infty(D_{-\delta/2}) \leq \exp\left(-\frac{\min_{i\in I}\{E(\mathcal{K}_i) + B(\mathcal{K}_i, D)\} - \min_{j\in I} E(\mathcal{K}_j)}{\eta} + \frac{c}{2\eta}\right). \tag{D.6.21}$$

But, for any $i \in I$, the exponent can be estimated as

$$\{E(\mathcal{K}_i) + B(\mathcal{K}_i, D)\} - \min_{j\in I} E(\mathcal{K}_i) \geq \begin{cases} E(\mathcal{K}_i) - \min_{j\in I} E(\mathcal{K}_j) & \text{if } i \notin J \\ B(\mathcal{K}_i, D) & \text{if } i \in J\,, \end{cases} \tag{D.6.22}$$

which is always positive, even in the first case, since $\arg\min_{j\in I} E(\mathcal{K}_j) \subset J$. Therefore, it holds that

$$\min_{i\in I}\{E(\mathcal{K}_i) + B(\mathcal{K}_i, D)\} - \min_{j\in I} E(\mathcal{K}_j) - \varepsilon \geq c - \varepsilon = \frac{c}{2}\,, \tag{D.6.23}$$

which concludes the proof. ∎

D.7. **Main results.** In this section, we restate the main results Theorems 1–4 and provide their proofs. They are now mostly corollaries of the results of Appendices D.5 and D.6.

Theorem D.1. *Suppose that μ_∞ is invariant under (SGD), fix a tolerance level $\varepsilon > 0$, and let $\mathcal{U}_i \equiv \mathcal{U}_i(\delta)$, $i = 1, \dots, K$, be δ-neighborhoods of the components of* crit f. *Then, for all sufficiently small $\delta, \eta > 0$, we have*

$$|\eta \log \mu_\infty(\mathcal{U}_i) + E_i - \min_j E_j| \leq \varepsilon \tag{D.7.1}$$

and

$$\left|\eta \log \frac{\mu_\infty(\mathcal{U}_i)}{\mu_\infty(\mathcal{U}_j)} + E_i - E_j\right| \leq \varepsilon. \tag{D.7.2}$$

More compactly, with notation as above, we have:

$$\mu_\infty(\mathcal{U}_i) \propto \exp\left(-\frac{E_i + \mathcal{O}(\varepsilon)}{\eta}\right). \tag{D.7.3}$$

Proof. Note that if μ_∞ is invariant for (SGD), it is *a fortiori* invariant for the accelerated process $(x_n^\eta)_{n\geq 0}$. This result is then a direct consequence of Proposition D.4 (Appendix D.5). ∎

Theorem D.2. *Suppose that μ_∞ is invariant under (SGD), and let $\mathcal{K}$ be a non-minimizing component of f. Then, with notation as in Theorem 1, there exists a minimizing component $\mathcal{K}'$ of f and a positive constant $c \equiv c(\mathcal{K}, \mathcal{K}') > 0$ such that*

$$\frac{\mu_\infty(\mathcal{U})}{\mu_\infty(\mathcal{U}')} \leq \exp\left(-\frac{c(\mathcal{K}, \mathcal{K}') + \varepsilon}{\eta}\right) \tag{D.7.4}$$

for all all sufficiently small $\eta > 0$ and all sufficiently small neighborhoods $\mathcal{U}$ and $\mathcal{U}'$ of $\mathcal{K}$ and $\mathcal{K}'$ respectively. In particular, in the limit $\eta \to 0$, we have $\mu_\infty(\mathcal{U}) \to 0$.

Proof. Let $\mathcal{K} = \mathcal{K}_i$ be a non-minimizing component. By Lemma D.32, there exists $j \in I$ such that $E_j < E_i$. The statement then follows from Theorem 1. ∎

Theorem D.3. *Suppose that μ_∞ is invariant under (SGD), fix a tolerance level $\delta > 0$, and let $\mathcal{U} \equiv \mathcal{U}(\delta)$ be a δ-neighborhood of* crit f. *Then there exists a constant $c \equiv c_\delta > 0$ such that, for all sufficiently small $\eta > 0$, we have:*

$$\mu_\infty(\mathcal{U}) \geq 1 - e^{-c/\eta}. \tag{D.7.5}$$

Proof. We actually show a slighly stronger result. Define $J \coloneqq \{i \in I : \mathcal{K}_i \text{ is minimizing}\}$. We prove that there exists $\mathcal{U}$ neighborhood of $\bigcup_{i\in J} \mathcal{K}_i$, a constant $c > 0$, such that, for all η small enough,

$$\mu_\infty\left(\mathcal{X} \setminus \bigcup_{i\in J} \mathcal{V}_i\right) \leq e^{-\frac{c}{\eta}}. \tag{D.7.6}$$

This is then a consequence of Proposition D.5 (Appendix D.6) with $J \leftarrow J$. Note J contain the ground states since they are minimizing by Lemma D.32. ∎

Theorem D.4. *Suppose that μ_∞ is invariant under (SGD), fix a tolerance level $\delta > 0$, and let $\mathcal{U}_0 \equiv \mathcal{U}_0(\delta)$ be a δ-neighborhood of the system's ground state $\mathcal{K}_0$. Then there exists a constant $c \equiv c_\delta > 0$ such that, for all sufficiently small $\eta > 0$, we have:*

$$\mu_\infty(\mathcal{U}_0) \geq 1 - e^{-c/\eta}. \tag{D.7.7}$$

Proof. It suffices to apply Proposition D.5 (Appendix D.6) with J as the set of ground states $\arg\min_{i\in I} E_i$. They are necessarily minimizing by Lemma D.32. ∎

D.8. **Extension: the mean occupation measures.** We now state and prove analogue of Theorems 1–4 for the mean occupation measure μ_n.

For this we will need a strengthened version of Assumption 3*, viz.

Assumption 3** (Variant of Assumption 3*)**.** The signal-to-noise ratio of G satisfies

$$\frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \to \infty \quad \text{as} \quad \|x\| \to \infty\,. \tag{4**}$$

In this section, we will posit that Assumption 3** holds in addition to Assumptions 1*, 2* and 4. We then have the following series of results for the occupation measure μ_n of (SGD).

Theorem 1* (Occupation variant of Theorem 1)**.** *Fix a tolerance level $\varepsilon > 0$, and let $\mathcal{U}_i \equiv \mathcal{U}_i(\delta)$, $i = 1, \dots, K$, be δ-neighborhoods of the components of* $\operatorname{crit} f$*. Then, for all sufficiently small $\delta, \eta > 0$ and large enough n, we have*

$$|\eta \log \mu_n(\mathcal{U}_i) + E_i - \min_j E_j| \leq \varepsilon \tag{D.8.1}$$

and

$$\left|\eta \log \frac{\mu_n(\mathcal{U}_i)}{\mu_n(\mathcal{U}_j)} + E_i - E_j\right| \leq \varepsilon. \tag{D.8.2}$$

More compactly, with notation as above, we have:

$$\mu_n(\mathcal{U}_i) \propto \exp\left(-\frac{E_i + \mathcal{O}(\varepsilon)}{\eta}\right). \tag{D.8.3}$$

Theorem 2* (Occupation variant of Theorem 2)**.** *Let $\mathcal{K}$ be a non-minimizing component of f. Then, with notation as in Theorem 1, there exists a minimizing component $\mathcal{K}'$ of f and a positive constant $c \equiv c(\mathcal{K}, \mathcal{K}') > 0$ such that*

$$\frac{\mu_n(\mathcal{U})}{\mu_n(\mathcal{U}')} \leq \exp\left(-\frac{c(\mathcal{K}, \mathcal{K}') + \varepsilon}{\eta}\right) \tag{D.8.4}$$

for all all sufficiently small $\eta > 0$, n large enough and all sufficiently small neighborhoods $\mathcal{U}$ and $\mathcal{U}'$ of $\mathcal{K}$ and $\mathcal{K}'$ respectively.

Theorem 3* (Occupation variant of Theorem 3)**.** *Fix a tolerance level $\delta > 0$, and let $\mathcal{U} \equiv \mathcal{U}(\delta)$ be a δ-neighborhood of* $\operatorname{crit} f$*. Then there exists a constant $c \equiv c_\delta > 0$ such that, for all sufficiently small $\eta > 0$ and large enough n, we have:*

$$\mu_n(\mathcal{U}) \geq 1 - e^{-c/\eta}. \tag{D.8.5}$$

Theorem 4* (Occupation variant of Theorem 4)**.** *Fix a tolerance level $\delta > 0$, and let $\mathcal{U}_0 \equiv \mathcal{U}_0(\delta)$ be a δ-neighborhood of the system's ground state $\mathcal{K}_0$. Then there exists a constant $c \equiv c_\delta > 0$ such that, for all sufficiently small $\eta > 0$ and large enough n, we have:*

$$\mu_n(\mathcal{U}_0) \geq 1 - e^{-c/\eta}. \tag{D.8.6}$$

We begin with a preliminary lemma which shows that, under Assumption 3**, the sequence of mean occupation measure $(\mu_n)_{n\geq 0}$ is tight (see e.g., Kallenberg [32, Chap. 23]).

Lemma D.34. *The sequence of mean occupation measures $(\mu_n)_{n\geq 0}$ is tight.*

The proof of this lemma first follows the proof of Lemma D.16 and then relies on the same reasoning as the proof of Douc et al. [14, Thm. 12.3.3].

Proof. By Lemma D.14, there exists $\mathcal{K} \subset \mathcal{X}$ compact, $\eta_0 > 0$, $c > 0$ such that, for any $\eta \leq \eta_0$, $x_0 = x \notin \mathcal{K}$,

$$U_\infty(x_1) - U_\infty(x) \leq \eta \left(\frac{\|\mathsf{U}(x, \omega_0)\|^2}{\sigma_\infty^2(f(x))} - \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \right). \tag{D.8.7}$$

Passing to the expectation yields that, for any $x \notin \mathcal{K}$,

$$\mathbb{E}_x[U_\infty(x_1)] - U_\infty(x) \leq \eta \left(\frac{\mathbb{E}_x\left[\|\mathsf{U}(x, \omega_0)\|^2\right]}{\sigma_\infty^2(f(x))} - \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \right). \tag{D.8.8}$$

Appplying Corollary D.2 with $X \leftarrow \frac{\mathsf{U}(x,\omega_0)}{\sqrt{\sigma_\infty^2(f(x))}}$ (the conditions of application are verified from Assumption 2*(c)) yields that

$$\mathbb{E}_x[U_\infty(x_1)] - U_\infty(x) \leq -\eta \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} + \eta \times 16 d \log 6 \,. \tag{D.8.9}$$

Hence, for any $x \in \mathcal{X}$, we have

$$\begin{aligned} \mathbb{E}_x[U_\infty(x_1)] - U_\infty(x) \leq{}& \mathbb{1}\{x \in \mathcal{K}\} \left(\sup_{x' \in \mathcal{K}} \mathbb{E}_{x'}[U_\infty(x')] - \inf_{\mathcal{X}} U_\infty \right) \\ & - \mathbb{1}\{x \notin \mathcal{K}\} \left(\eta \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} - \eta \times 16 d \log 6 \right), \end{aligned} \tag{D.8.10}$$

or, after rearranging,

$$\begin{aligned} \mathbb{E}_x[U_\infty(x_1)] & - \mathbb{1}\{x \in \mathcal{K}\} \left(\sup_{x' \in \mathcal{K}} \mathbb{E}_{x'}[U_\infty(x')] - \inf_{\mathcal{X}} U_\infty \right) \\ & + \mathbb{1}\{x \notin \mathcal{K}\} \left(\eta \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} - \eta \times 16 d \log 6 \right) \leq U_\infty(x) \,. \end{aligned} \tag{D.8.11}$$

Since the function

$$\begin{aligned} x \mapsto & - \mathbb{1}\{x \in \mathcal{K}\} \left(\sup_{x' \in \mathcal{K}} \mathbb{E}_{x'}[U_\infty(x')] - \inf_{\mathcal{X}} U_\infty \right) \\ & + \mathbb{1}\{x \notin \mathcal{K}\} \left(\eta \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} - \eta \times 16 d \log 6 \right) \end{aligned} \tag{D.8.12}$$

is measurable, lower-bounded and goes to infinity as $\|x\| \to \infty$ by Assumption 3**, one can then apply the same computations as in the proof of Douc et al. [14, Thm. 12.3.3] to obtain that the sequence of occupation measures $(\mu_n)_{n \geq 0}$ is tight. ∎

We now prove Theorem 1* by adapting the proof of Theorem 1. Since the process is exactly the same for Theorems 2*–4*, we omit their proofs.

Proof of Theorem 1.* We show that, for sufficiently small $\delta, \eta > 0$, for any $i \in I$, we have that

$$\begin{aligned} \exp\left(-\frac{E_i - \min_j E_j + \varepsilon}{\eta} \right) & \leq \liminf_{n \to \infty} \mu_n(\mathcal{U}_i) \\ & \leq \limsup_{n \to \infty} \mu_n(\mathcal{U}_i) \\ & \leq \exp\left(-\frac{E_i - \min_j E_j - \varepsilon}{\eta} \right) \end{aligned} \tag{D.8.13}$$

and the results in the statement will follow with 2ε in place of ε.

We apply Proposition D.4 (Appendix D.5): take δ small enough so Proposition D.4 can be applied with both the neighborhoods $\mathcal{U}_1, \dots, \mathcal{U}_K$ and $\mathrm{cl}\,\mathcal{U}_1, \dots, \mathrm{cl}\,\mathcal{U}_K$. One then obtain

$\eta_0 > 0$ such that, for all $0 < \eta < \eta_0$ and any μ_∞ invariant probability measure for $(x_n^\eta)_n$, for any $i \in I$,

$$\begin{aligned} \mu_\infty(\mathcal{U}_i) &\geq \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j\in I} E(\mathcal{K}_j)}{\eta} - \frac{\varepsilon}{\eta}\right) \\ \mu_\infty(\mathrm{cl}\,\mathcal{U}_i) &\leq \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j\in I} E(\mathcal{K}_j)}{\eta} + \frac{\varepsilon}{\eta}\right). \end{aligned} \tag{D.8.14}$$

We now prove that Eq. (D.8.13) holds. Fix $i \in I$. By Lemma D.34, the sequence of mean occupation measures $(\mu_n)_{n\geq 0}$ is tight, so that, by Prohorov theorem [32, Thm. 23.2], it is sequentially compact for the weak topology, or, in other terms, for the convergence in distribution. Therefore, $(\mu_n)_{n\geq 0}$ admits a weak accumulation point which is a probability distribution and that we denote by ν. Applying Portmanteau theorem [32, Thm. 5.25] to the open set $\mathcal{U}_i$ and the closed set $\mathrm{cl}\,\mathcal{U}_i$ yields that

$$\nu(\mathcal{U}_i) \leq \liminf_{n\to\infty} \mu_n(\mathcal{U}_i) \leq \limsup_{n\to\infty} \mu_n(\mathrm{cl}\,\mathcal{U}_i) \leq \nu(\mathrm{cl}\,\mathcal{U}_i)\,. \tag{D.8.15}$$

Since $(x_n)_{n\geq 0}$, the sequence of iterates of SGD, is (weak) Feller by Lemma D.15, ν is actually invariant for $(x_n)_{n\geq 0}$, by, e.g., Douc et al. [14, Prop. 12.3.1], and *a fortiori* invariant for the accelerated process $(x_n^\eta)_{n\geq 0}$. Combining Eq. (D.8.14) with Eq. (D.8.15) gives the result Eq. (D.8.13). ∎

Appendix E. Potential for the invariant measure

E.1. **Gaussian noise.** Though it does not formally fit into our setting, let us first begin with the case where the noise is Gaussian. Since it is unbounded, our assumptions are not satisfied and our theorems describing the invariant measure do not apply. However, all the objects we consider are still well-defined, and, in that case, it is possible to compute the E_i explicitely. Moreover, this section serves as a blueprint for the truncated Gaussian case of the next section.

Assume that, for every $x \in \mathcal{X}$, $\mathsf{U}(x,\omega)$ follows a centered Gaussian distribution with covariance $\sigma^2(f(x))I$ for some continuous function $\sigma^2 : \mathbb{R} \to (0, +\infty)$.

Akin to U_∞ in Appendix D, a key role is played by the function $U : \mathcal{X} \to \mathbb{R}$ defined by

$$U(x) \coloneqq 2\alpha(f(x)) \quad \text{with} \quad \alpha\colon \mathbb{R} \to \mathbb{R} \quad \text{a primitive of} \quad 1/\sigma^2\,. \tag{E.1.1}$$

Since the noise is Gaussian, the Lagrangian and Hamiltonian have explicit expressions: for every $x, p, v \in \mathcal{X}$,

$$\mathcal{H}(x,p) = -\langle \nabla f(x), p\rangle + \frac{1}{2}\sigma^2(f(x))\|p\|^2 \tag{E.1.2a}$$

$$\mathcal{L}(x,v) = \frac{\|v + \nabla f(x)\|}{2\sigma^2(f(x))}\,. \tag{E.1.2b}$$

This expression of $\mathcal{L}$ make it clear that the action function penalizes the deviation of a path from the flow: for a path $\gamma \in \mathcal{C}([0,T])$ for some $T > 0$,

$$\mathcal{S}_T(\gamma) = \int_0^T \frac{\|\dot\gamma_t + \nabla f(\gamma_t)\|^2}{2\sigma^2(f(\gamma_t))}\,dt\,. \tag{E.1.3}$$

The computation of the E_i relies on the following observation. Take a path $\gamma \in \mathcal{C}([0,T])$ for some $T > 0$ and consider φ defined by $\varphi_t = \gamma_{T-t}$ for $t \in [0,T]$. Then, the action cost of φ is given by,

$$\mathcal{S}_T(\varphi) = \int_0^T \frac{\|-\dot\gamma_t + \nabla f(\gamma_t)\|^2}{2\sigma^2(f(\gamma_t))}\,dt$$

$$= \int_0^T \frac{\|\dot\gamma_t + \nabla f(\gamma_t)\|^2}{2\sigma^2(f(\gamma_t))}\,dt - \int_0^T \frac{2\langle \dot\gamma_t, \nabla f(\gamma_t)\rangle}{\sigma^2(f(\gamma_t))}\,dt$$
$$= \mathcal{S}_T(\gamma) - \int_0^T \langle \dot\gamma_t, \nabla U(\gamma_t)\rangle\,dt\,, \tag{E.1.4}$$

since $\nabla U(x) = 2\nabla\alpha(f(x))\nabla f(x)$. Therefore, we get that

$$\mathcal{S}_T(\varphi) = \mathcal{S}_T(\gamma) - (U(\gamma_T) - U(\gamma_0))\,. \tag{E.1.5}$$

Take $i, j \in I$. This equality then translates to a relation between $B_{i,j}$ and $B_{j,i}$: considering $\gamma \in \mathcal{C}([0,T])$ such that $\gamma_0 \in \mathcal{K}_i$, $\gamma_T \in \mathcal{K}_j$ and taking the infimum over all such paths, we get that

$$B_{j,i} \le B_{i,j} + (U_j - U_i)\,. \tag{E.1.6}$$

where, since f is constant on $\mathcal{K}_i$ and $\mathcal{K}_j$, we denote by U_i and U_j the values of U on $\mathcal{K}_i$ and $\mathcal{K}_j$ respectively.

Reversing the roles of i and j and applying the same argument shows that this inequality is an equality:

$$B_{j,i} + U_j = B_{i,j} + U_i\,. \tag{E.1.7}$$

Denote by $C_{i,j}$ this common value. Crucially, $C_{i,j}$ is symmetric in i and j.

Consider now T^* a minimum weight spanning tree on the complete but now undirected graph on I with weights $(C_{i,j})_{i,j}$. We show that the minima in the E_i are attained T^*, or more precisely, a directed version of it.

Fix i and consider $T \in \mathcal{T}_i$ a spanning tree rooted at i. We have that, since any node j is the origin of exactly one edge in T,

$$\sum_{(j\to l)\in T} B_{j,l} + \sum_{j\in I} U_j = \sum_{(j\to l)\in T} (B_{j,l} + U_j) + U_i$$
$$= \sum_{(j\to l)\in T} C_{j,l} + U_i\,. \tag{E.1.8}$$

But by definition of T^*, this sum is at least greater than $\sum_{(j\leftrightarrow l)\in T^*} C_{j,l}$ so that we have

$$\sum_{(j\to l)\in T} B_{j,l} + \sum_{j\in I} U_j \ge \sum_{(j\leftrightarrow l)\in T^*} C_{j,l} + U_i\,, \tag{E.1.9}$$

and taking T_i an oriented version of T^* rooted at i, the equality is attained. Therefore, we have that

$$E_i = \sum_{(j\to l)\in T_i} B_{j,l} - \sum_{j\in I} U_j + U_i\,, \tag{E.1.10}$$

or, in short, $E_i = U_i + c$ where $c > 0$ is independent of i.

Therefore, the mass distribution over critical points is governed by a Gibbs measure with potential U.

Let us now mention two particular cases.

- If σ^2 is constant, then $U = \frac{2f}{\sigma^2}$.
- It σ^2 is linear, i.e., of the form $\sigma^2(f(x)) = \sigma_1^2(f(x) + \sigma_0^2)$, then $U = \frac{2}{\sigma_1^2}\log(f + \sigma_0^2)$.

E.2. **Truncated Gaussian noise.** To fit into our theoretical framework, we consider truncated Gaussian noise instead. The general outline of the proof but with added steps to handle the error due to the truncation. In particular, one must show that, without loss of generality, we can only conider paths whose derivative has the same norm as the gradient of f. This is done with Freidlin & Wentzell [20, Chap. 4, Lem. 3.1] that we adapt to our setting.

Assume that $\mathsf{U}(x,\omega)$ follows a centered Gaussian distribution with covariance $\sigma^2(f(x))I$ conditioned on being in $\mathbb{B}(0,R(x))$ for some $R(x)>0$.

As in Definition 4, we define $U(x)=2\alpha(f(x))$ with $\alpha'=\frac{1}{\sigma^2}$ and denote by U_i the value taken by U on $\mathcal{K}_i$.

Consider some $0<\delta\leq\frac{1}{2}$ and assume that

$$\sup_{x\in\mathcal{X}} 2^{d+4}(d+1)e^{-\frac{R^2}{16\sigma^2}} \leq \delta\,, \tag{E.2.1}$$

so that the error term $2E(\sigma^2(f(x)),R(x))$ in Lemma F.2 is bounded by δ.

Moreover, assume that, for any $x\in\mathcal{X}$

$$\|\nabla f(x)\| \leq \frac{R(x)}{8} \tag{E.2.2}$$

Lemma E.1. *Consider $\gamma\in\mathcal{C}([0,T])$. Then, there exists $\widetilde{\gamma}\in\mathcal{C}([0,S])$ a reparametrization of γ such that, for any $t\in[0,S]$,*

$$\|\dot{\widetilde{\gamma}}_s\| = \|\nabla f(\widetilde{\gamma}_s)\|\,. \tag{E.2.3}$$

and

$$\mathcal{S}_T(\gamma) \geq \int_0^S \frac{\|\dot{\widetilde{\gamma}}_s+\nabla f(\widetilde{\gamma}_s)\|^2}{2(1+\delta)\sigma^2(f(\widetilde{\gamma}_s))}\,ds\,. \tag{E.2.4}$$

Proof. By the proof Freidlin & Wentzell [20, Chap. 4, Lem. 3.1], there exists $t(s)$ change of time such that, with $\widetilde{\gamma}_s=\gamma_{t(s)}$, $\|\dot{\widetilde{\gamma}}_s\|=\|\nabla f(\widetilde{\gamma}_s)\|$.

We have that

$$\mathcal{S}_T(\gamma) = \int_0^{t^{-1}(T)} \dot{t}(s)\mathcal{L}(\widetilde{\gamma}_s,(\dot{t}(s))^{-1}\dot{\widetilde{\gamma}}_s)\,ds\,, \tag{E.2.5}$$

so it suffices to bound $\mathcal{L}(\widetilde{\gamma}_s,\dot{\widetilde{\gamma}}_s)$ from below: by definition, we have

$$\begin{aligned}\mathcal{L}(\widetilde{\gamma}_s,(\dot{t}(s))^{-1}\dot{\widetilde{\gamma}}_s) &\geq \sup\left\{\langle p,(\dot{t}(s))^{-1}\dot{\widetilde{\gamma}}_t+\nabla f(\widetilde{\gamma}_s)\rangle - \bar{\mathcal{H}}(\widetilde{\gamma}_s,p) : \|p\|\leq\frac{R(\widetilde{\gamma}_s)}{2\sigma^2(f(\widetilde{\gamma}_s))}\right\}\\ &\geq \sup\left\{\langle p,(\dot{t}(s))^{-1}\dot{\widetilde{\gamma}}_t+\nabla f(\widetilde{\gamma}_s)\rangle - (1+\delta)\frac{\sigma^2(f(\widetilde{\gamma}_s))}{2}\|p\|^2 : \|p\|\leq\frac{R(\widetilde{\gamma}_s)}{2\sigma^2(f(\widetilde{\gamma}_s))}\right\},\end{aligned} \tag{E.2.6}$$

by Lemma F.2. Applying Lemma F.3 with $\lambda\leftarrow\dot{t}(s)$, now exactly yields, for almost all s,

$$\begin{aligned}\mathcal{L}(\widetilde{\gamma}_s,(\dot{t}(s))^{-1}\dot{\widetilde{\gamma}}_s) &\geq \dot{t}(s)\sup\left\{\langle p,\dot{\widetilde{\gamma}}_t+\nabla f(\widetilde{\gamma}_s)\rangle - (1+\delta)\frac{\sigma^2(f(\widetilde{\gamma}_s))}{2}\|p\|^2 : \|p\|\leq\frac{R(\widetilde{\gamma}_s)}{2\sigma^2(f(\widetilde{\gamma}_s))}\right\}\\ &= \dot{t}(s)\frac{\|\dot{\widetilde{\gamma}}_s+\nabla f(\widetilde{\gamma}_s)\|^2}{2(1+\delta)\sigma^2(f(\widetilde{\gamma}_s))}\,,\end{aligned} \tag{E.2.7}$$

since $\|\dot{\widetilde{\gamma}}_s+\nabla f(\widetilde{\gamma}_s)\|\leq\frac{R(\widetilde{\gamma}_s)}{2}$. ∎

Lemma E.2. *With this setting, for any $i\in I$*

$$E_i = U_i + \sum_{(j\to l)\in T_i} B_{j,l} - \sum_{j\in I} U_j + \mathcal{O}(\delta) \tag{E.2.8}$$

Proof. Consider $\gamma\in\mathcal{C}([0,T])$ such that $\gamma_0\in\mathcal{K}_i$, $\gamma_T\in\mathcal{K}_j$ and $\mathcal{S}_T(\gamma)<+\infty$. Then, by the previous lemma Lemma E.1, there exists $\widetilde{\gamma}\in\mathcal{C}([0,S])$ a reparametrization of γ such that, for any $t\in[0,S]$,

$$\mathcal{S}_T(\gamma) \geq \int_0^S \frac{\|\dot{\widetilde{\gamma}}_s+\nabla f(\widetilde{\gamma}_s)\|^2}{2(1+\delta)\sigma^2(f(\widetilde{\gamma}_s))}\,ds$$

$$= \int_0^S \frac{\|-\dot{\widetilde{\gamma}}_s + \nabla f(\widetilde{\gamma}_s)\|^2}{2(1+\delta)\sigma^2(f(\widetilde{\gamma}_s))}\, ds + \frac{U(\gamma_T) - U(\gamma_0)}{1+\delta}\,, \tag{E.2.9}$$

where we performed the same computations as above in Appendix E.1. Considering the path $(\widetilde{\gamma}_{S-s})_{s\in[0,S]}$ and invoking the upper-bound on the Lagrangian from Lemma F.2 with $-\dot{\widetilde{\gamma}}_s + \nabla f(\widetilde{\gamma}_s)$ which still has norm less than $R(\widetilde{\gamma}_s)/4$, we get that

$$\mathcal{S}_T(\gamma) \geq \frac{1-\delta}{1+\delta} B_{j,i} + \frac{U(\gamma_T) - U(\gamma_0)}{1+\delta}\,. \tag{E.2.10}$$

The result now follows from the same computations as in Appendix E.1. ∎

E.3. **Local dependencies.** Under local assumptions similar to Mori et al. [56], we demonstrate how the modelling of the noise influences the invariant measure.

Lemma E.3. *Consider $\mathcal{K}_i$ minimizing, $\sigma^2\colon \mathbb{R} \to (0,+\infty)$ continuous, and take $\alpha\colon \mathbb{R}\to\mathbb{R}$ such that $\alpha' = \frac{1}{\sigma^2}$. Assume that H^* is a positive definite matrix such that, locally near $\mathcal{K}_i$, it holds that:*

$$\alpha(f(x)) = \sum_{\lambda\in \operatorname{eig} H^*} g^\lambda(x_\lambda)\,, \tag{E.3.1}$$

where x_λ denotes the orthogonal projection of x on the eigenspace of the eigenvalue λ and where $g^\lambda\colon \mathcal{X}\to\mathbb{R}$ is continuously differentiable. Define the potential $U\colon \mathcal{X}\to\mathbb{R}$ by

$$U(x) = \sum_{\lambda\in \operatorname{eig} H^*} \frac{2g^\lambda(x_\lambda)}{\lambda}\,. \tag{E.3.2}$$

If we have the anistropic subGaussian bound: for x in a neighborhood of $\mathcal{K}_i$, $p\in\overline{\mathbb{B}}(0,\|U(x)\|)$,

$$\bar{\mathcal{H}}(x,p) \leq \frac{\sigma^2(f(x))}{2}\langle p, H^* p\rangle\,, \tag{E.3.3}$$

then there is $\delta>0$ such that, for all $j\neq i$, any $0<\delta'\leq\delta$,

$$E_j \geq \min\{U(x) - U_i : x, d(x,\mathcal{K}_i) = \delta'\} > 0\,, \tag{E.3.4}$$

where U_i is the value of U on $\mathcal{K}_i$.

Moreover, there exists $R>\delta$ such that, if there exists $c>0$, $s^2>0$, such that, for all $x\in\overline{\mathbb{B}}(0,R)\setminus\mathcal{U}_\delta(\mathcal{K}_i)$, $v\in\mathbb{B}(\nabla f(x),c)$,

$$\bar{\mathcal{L}}(x,v) \leq \frac{\|v\|}{2s^2}\,, \tag{E.3.5}$$

then there exists $C>0$ that depends only on r,R,c, f restricted to $\mathcal{X}\setminus\mathcal{U}_r(\mathcal{K}_i)$ such that

$$E_i \leq \frac{C}{s^2}\,. \tag{E.3.6}$$

The assumptions on the Hamiltonian and the Lagrangian roughly say that the noise share some similarities with Gaussian distributions with the prescribed variances. In particular, they are satisfied in the (truncated) Gaussian case, as shown in Lemma F.2.

Moreover, a takeaway of this lemma is that, if σ^2 or the eigenvalues of H^* are small enough, then $\mathcal{K}_i$ must be the ground state even if it may not be the global minimum of U. This lemma is general enough to handle non-constant variance: a notable example is when $\sigma^2(f(x))$ is linear in $f(x)$ and where the resulting potential U then depends logarithmically on the value f.

Proof. Denote $P_\lambda \in \mathbb{R}^{d\times d}$ the orthogonal projection on the eigenspace of H^* associated with the eigenvalue λ. U can thus be rewritten as

$$.U(x) = \sum_{\lambda\in\operatorname{eig} H^*} \frac{2g^\lambda(P_\lambda x)}{\lambda}\,, \tag{E.3.7}$$

so that its gradient is given by

$$\nabla U(x) = \sum_{\lambda\in\operatorname{eig} H^*} \frac{2P_\lambda \nabla g^\lambda(P_\lambda x)}{\lambda}\,. \tag{E.3.8}$$

In particular, we obtain that for x close enough to $\mathcal{K}_i$, using the orthogonality of the projections,

$$\begin{aligned}
\bar{\mathcal{H}}(x,\nabla U(x)) &\le \frac{\sigma^2(f(x))}{2}\langle \nabla U(x), H^*\nabla U(x)\rangle \\
&= \frac{\sigma^2(f(x))}{2}\sum_{\lambda\in\operatorname{eig} H^*}\frac{4\|P_\lambda\nabla g^\lambda(P_\lambda x)\|^2}{\lambda} \\
&= \frac{\sigma^2(f(x))}{2}\langle \nabla U(x), 2\nabla(\alpha\circ f)(x)\rangle \\
&= \langle \nabla U(x), \nabla f(x)\rangle\,.
\end{aligned} \tag{E.3.9}$$

Therefore, for x close enough to $\mathcal{K}_i$, we have that

$$\mathcal{H}(x,\nabla U(x)) = -\langle \nabla U(x),\nabla f(x)\rangle + \bar{\mathcal{H}}(x,\nabla U(x)) \le 0\,. \tag{E.3.10}$$

For x close to $\mathcal{K}_i$, let us compute $\langle \nabla U(x),\nabla f(x)\rangle$:

$$\begin{aligned}
\langle \nabla U(x),\nabla f(x)\rangle &= \sigma^2(f(x))\langle \nabla U(x), \nabla(\alpha\circ f)(x)\rangle \\
&= \sigma^2(f(x))\sum_{\lambda\in\operatorname{eig} H^*}\frac{2\|P_\lambda\nabla g^\lambda(P_\lambda x\|}{\lambda}\,,
\end{aligned} \tag{E.3.11}$$

which is (stricly) positive in a small neighborhood of $\mathcal{K}_i$ (excluding $\mathcal{K}_i$ itself). Therefore, U is decreasing on trajectories of the flow in this neighborhood. With the same proof as in Lemma D.29, we deduce that there exists $\delta > 0$ such that $\arg\min_{x\in\mathcal{U}_\delta(\mathcal{K}_i)} U(x) = \mathcal{K}_i$. Moreover, take $\delta > 0$ small enough such that Eq. (E.3.10) holds on $\mathcal{U}_\delta(\mathcal{K}_i)$, $\mathcal{U}_\delta(\mathcal{K}_i)\cap \operatorname{crit} f = \mathcal{K}_i$ and the trajectories of the flow started in $\mathcal{U}_\delta(\mathcal{K}_i)$ stay converge to $\mathcal{K}_i$. We now proceed as in the proof of Lemma D.33. Take $\delta' \le \delta/2$, $\mathcal{U} := \mathcal{U}'_\delta(\mathcal{K}_i)$ and $\Delta := \min\{U(x) - U_i : x\in\mathcal{X},\, d(x,\mathcal{K}_i) = \delta'\}$, which is positive by definition. Fix $j\neq i$: we show that $B_{i,j}\ge\Delta$. Consider some $T>0$ and $\gamma\in\mathcal{C}([0,T],\mathcal{X})$ such that $\gamma_0\in\mathcal{K}_i$ and $\gamma_T\in\mathcal{K}_j$. By definition of δ and by continuity of γ and $d(\cdot,\mathcal{K}_i)$, there exists $t\in[0,T]$ such that $d(\gamma_t,\mathcal{K}_i) = \delta'$. Therefore, we have that

$$\mathcal{S}_{0,T}(\gamma) \ge \mathcal{S}_{0,t}(\gamma) = \int_0^t \mathcal{L}(\gamma_s,\dot\gamma_s)\,ds\,, \tag{E.3.12}$$

and therefore, by definition of $\mathcal{L}$ as the conjugate of $\mathcal{H}$, we obtain that

$$\mathcal{S}_{0,T}(\gamma) \ge \int_0^t \langle\dot\gamma_s,\nabla U(\gamma_s)\rangle - \bar{\mathcal{H}}(\gamma_s,\nabla U(\gamma_s))\,ds \ge \int_0^t \langle\dot\gamma_s,\nabla U(\gamma_s)\rangle\,, \tag{E.3.13}$$

where we used Eq. (E.3.10) in the last inequality. Thus, we get

$$\mathcal{S}_{0,T}(\gamma) \ge U(\gamma_t) - U(\gamma_0) \ge \Delta\,, \tag{E.3.14}$$

with any $\delta'\le\delta/2$ (and therefore δ in the statement corresponds to $\delta/2$ here).

Finally, it remains to show that, any $l \neq i$, $E_l \geq \Delta$. Consider any tree rooted at l: it must have an edge of the form $i \to j$ for some $j \in I$ and therefore the sum of its weights will be at least $B_{i,j} \geq \Delta$. Hence, since it holds for any such tree, we have that $E_l \geq \Delta$.

We now prove the second part of the lemma. δ was chosen small enough so that it is possible to find $R > \delta$ such that $\mathbb{B}(0,R) \setminus \mathcal{U}_\delta(\mathcal{K}_i)$ contains all the $\mathcal{K}_j$ for $j \neq i$ and not $\mathcal{K}_i$. The assumption on the Lagrangian implies that, for any $x \in \mathbb{B}(0,R) \setminus \mathcal{U}_\delta(\mathcal{K}_i)$, $v \in \mathbb{B}(0,c)$,

$$\mathcal{L}(x,v) \leq \frac{\|v + \nabla f(x)\|^2}{2s^2} \,. \tag{E.3.15}$$

Fix $j \neq i$ and take $\gamma \in \mathcal{C}^1([0,T],\mathcal{X})$ such that $\gamma_0 \in \mathcal{K}_j$, $x := \gamma_T \in \mathcal{U}_\delta(\mathcal{K}_i)$ and γ remains in $\overline{\mathbb{B}}(0,R) \setminus \mathcal{U}_\delta(\mathcal{K}_i)$. Without loss of generality, at the expense of replacing γ by a reparametrization, we can assume that $\dot\gamma_t$ in $\mathbb{B}(0,c)$ for all $t \in [0,T]$. Then, we have that

$$\mathcal{S}_{0,T}(\gamma) \leq \int_0^T \mathcal{L}(\gamma_s, \dot\gamma_s)\, ds \leq \int_0^T \frac{\|\dot\gamma_s + \nabla f(\gamma_s)\|^2}{2s^2}\, ds \,, \tag{E.3.16}$$

which depends only on the value of ∇f outside of $\mathcal{U}_{\delta/2}(\mathcal{K}_i)$. But, by the same reasoning as in Lemma D.31, since the flow started at x converges to $\mathcal{K}_i$, we have that $B(\{x\},\mathcal{K}_i) = 0$. Therefore, we have that,

$$B_{j,i} \leq \int_0^T \frac{\|\dot\gamma_s + \nabla f(\gamma_s)\|^2}{2s^2}\, ds \,, \tag{E.3.17}$$

and, taking the maximum of such quantities over all $j \neq i$, we obtain $C > 0$ such, for any $j \neq i$,

$$B_{j,i} \leq \frac{C}{s^2} \,. \tag{E.3.18}$$

To conclude on the value of E_i, we consider the tree rooted at i made of all the edges (j,i) for $j \neq i$. It has weight at most $(K-1)C/s^2$ and therefore, we have that $E_i \leq (K-1)C/s^2$. ∎

Appendix F. Auxiliary results

F.1. **Truncated Gaussian distribution.**

Lemma F.1. *Consider $X \sim \mathcal{N}(0,\Sigma)$ a multivariate Gaussian distribution $\Sigma \in \mathbb{R}^{d\times d}$ positive definite. For $R > 0$, define the truncated Gaussian random variable (r.v.) X_R by conditioning X to the ball $\overline{\mathbb{B}}(0,R)$. Define*

$$\bar{\mathcal{H}}(p) := \log \mathbb{E}\Big[e^{\langle p, X_R\rangle}\Big] \tag{F.1.1}$$

and

$$\widetilde{E}(\Sigma,\widetilde{R}) := e^{-\frac{\widetilde{R}^2}{4\|\Sigma\|}}(\operatorname{tr}\Sigma + \|\Sigma\|)2^{d+3} \tag{F.1.2}$$

Then, for $\widetilde{R} > 0$ such that

$$\widetilde{R} \geq \sqrt{\|\Sigma\|(2d+4)\log 2} \,, \tag{F.1.3}$$

it holds that, for any $p \in \mathcal{X}$ such that $\|\Sigma p\| \leq R - \widetilde{R}$,

$$\big|\bar{\mathcal{H}}(p) - \tfrac{1}{2}\langle p, \Sigma p\rangle\big| \leq \tfrac{1}{2}\widetilde{E}(\Sigma,\widetilde{R})\|p\|^2 \tag{F.1.4a}$$

$$\big\|\nabla\bar{\mathcal{H}}(p) - \Sigma p\big\| \leq \widetilde{E}(\Sigma,\widetilde{R})\|p\| \tag{F.1.4b}$$

$$\big\|\operatorname{Hess}\bar{\mathcal{H}}(p) - \Sigma\big\| \leq \widetilde{E}(\Sigma,\widetilde{R}) \,. \tag{F.1.4c}$$

Proof. Define, for S a measurable set,

$$Z(S) = \int_S e^{-\frac{1}{2}x\cdot\Sigma^{-1}x}\, dx \,, \tag{F.1.5}$$

and, for convenience, $\mathcal{K} \coloneqq \overline{\mathbb{B}}(0,R)$. For notational convenience, in this proof, we will denote the inner product between two vectors x and p with a simple dot $x \cdot p$. We have that

$$\begin{aligned}\mathbb{E}\big[e^{X_R \cdot p}\big] &= \frac{e^{\frac{1}{2}p\cdot\Sigma p}}{Z(\mathcal{K})}\int_{\mathcal{K}} e^{-\frac{1}{2}(x-\Sigma p)\cdot\Sigma^{-1}(x-\Sigma p)}\,dx \\ &= \frac{e^{\frac{1}{2}p\cdot\Sigma p}}{Z(\mathcal{K})}\int_{\mathcal{K}-\Sigma p} e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' \\ &= e^{\frac{1}{2}p\cdot\Sigma p}\frac{Z(\mathcal{K}-\Sigma p)}{Z(\mathcal{K})}\,, \end{aligned} \tag{F.1.6}$$

where we performed the change of variable $x' \leftarrow x - \Sigma p$.

Define

$$f(p) \coloneqq Z(\mathcal{K}-\Sigma p) = \int_{\mathcal{K}} e^{-\frac{1}{2}(x-\Sigma p)\cdot\Sigma^{-1}(x-\Sigma p)}\,dx\,. \tag{F.1.7}$$

so that

$$\bar{\mathcal{H}}(p) = \log\mathbb{E}\big[e^{X_R\cdot p}\big] = \frac{1}{2}p\cdot\Sigma p + \log\frac{f(p)}{f(0)}\,. \tag{F.1.8}$$

Therefore it suffices to bound $\log\frac{f(p)}{f(0)}$ and its derivatives.

Differentiating yields

$$\nabla f(p) = \int_{\mathcal{K}} (x-\Sigma p)e^{-\frac{1}{2}(x-\Sigma p)\cdot\Sigma^{-1}(x-\Sigma p)}\,dx \tag{F.1.9a}$$

$$\operatorname{Hess} f(p) = \int_{\mathcal{K}} (x-\Sigma p)(x-\Sigma p)^\top e^{-\frac{1}{2}(x-\Sigma p)\cdot\Sigma^{-1}(x-\Sigma p)}\,dx - \Sigma f(p)\,. \tag{F.1.9b}$$

Note that, by symmetry of $\overline{\mathbb{B}}(0,R)$, $\nabla f(0) = 0$.

We first bound $\operatorname{Hess} f(p)$. Performing the change of variable $x' \leftarrow x - \Sigma p$ again gives us

$$\begin{aligned}\operatorname{Hess} f(p) &= \int_{\mathcal{K}-\Sigma p} x'x'^\top e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' - \Sigma f(p) \\ &= \int_{\mathcal{K}-\Sigma p} (x'x'^\top - \Sigma)e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' \\ &= -\int_{\mathcal{X}\setminus(\mathcal{K}-\Sigma p)} (x'x'^\top - \Sigma)e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx'\,, \end{aligned} \tag{F.1.10}$$

where we used that $\int_{\mathcal{X}}(x'x'^\top - \Sigma)e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' = 0$.

By definition of $\widetilde{R}$, $\mathcal{K}-\Sigma p$ contains $\overline{\mathbb{B}}(0,\widetilde{R})$. We now bound the operator norm of $\operatorname{Hess} f(p)$:

$$\begin{aligned}\|\operatorname{Hess} f(p)\| &\le \int_{\mathcal{X}\setminus(\mathcal{K}-\Sigma p)} (\|x'\|^2 + \|\Sigma\|)e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' \\ &\le \int_{\mathcal{X}\setminus\overline{\mathbb{B}}(0,\widetilde{R})} (\|x'\|^2 + \|\Sigma\|)e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' \\ &\le e^{-\frac{\widetilde{R}^2}{4\|\Sigma\|}}\int_{\mathcal{X}} (\|x'\|^2 + \|\Sigma\|)e^{-\frac{1}{4}x'\cdot\Sigma^{-1}x'}\,dx' \\ &= e^{-\frac{\widetilde{R}^2}{4\|\Sigma\|}}(\operatorname{tr}\Sigma + \|\Sigma\|)(4\pi)^{d/2}\det(\Sigma)^{1/2}\,, \end{aligned} \tag{F.1.11}$$

We now bound $\nabla f(p)$. The change of variable $x' \leftarrow x - \Sigma p$ yields

$$\nabla f(p) = \int_{\mathcal{K}-\Sigma p} x' e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' = \int_{\mathcal{X}\setminus(\mathcal{K}-\Sigma p)} x' e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx'\,, \tag{F.1.12}$$

where we used that $\int_{\mathcal{X}} x' e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' = 0$.

Therefore, similar computations as above yield that

$$\begin{aligned}
\|\nabla f(p)\| &\le \int_{\mathcal{X}\setminus(\mathcal{K}-\Sigma p)} \|x'\| e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' \\
&\le \int_{\mathcal{X}\setminus\overline{\mathbb{B}}(0,\widetilde{R})} \|x'\| e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' \\
&\le e^{-\frac{\widetilde{R}^2}{4\|\Sigma\|}} \int_{\mathcal{X}} \|x'\| e^{-\frac{1}{4}x'\cdot\Sigma^{-1}x'}\,dx' \\
&\le e^{-\frac{\widetilde{R}^2}{4\|\Sigma\|}} \sqrt{\operatorname{tr}\Sigma}(4\pi)^{d/2}\det(\Sigma)^{1/2}\,.
\end{aligned} \tag{F.1.13}$$

Let us now lower-bound $f(p)$. In the same fashion as above, we have that

$$\begin{aligned}
f(p) &= \int_{\mathcal{K}-\Sigma p} e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' \\
&\ge \int_{\overline{\mathbb{B}}(0,\widetilde{R})} e^{-\frac{1}{2}x'\cdot\Sigma^{-1}x'}\,dx' \\
&\ge (2\pi)^{d/2}\det(\Sigma)^{1/2}\left(1 - 2^{d/2} e^{-\frac{\widetilde{R}^2}{4\|\Sigma\|}}\right),
\end{aligned} \tag{F.1.14}$$

so that, if $\widetilde{R} \ge \sqrt{\|\Sigma\|(2d+4)\log 2}$, it holds that

$$f(p) \ge \tfrac{1}{2}(2\pi)^{d/2}\det(\Sigma)^{1/2} \tag{F.1.15}$$

The Hessian of $\log f/f(0)$ is then given by

$$\operatorname{Hess}\log\frac{f(p)}{f(0)} = \frac{\operatorname{Hess} f(p)}{f(p)} - \frac{\nabla f(p)\nabla f(p)^\top}{f(p)^2}\,. \tag{F.1.16}$$

Eqs. (F.1.11), (F.1.13) and (F.1.15) combined yield that

$$\begin{aligned}
\left\|\operatorname{Hess}\log\frac{f(p)}{f(0)}\right\| &\le e^{-\frac{\widetilde{R}^2}{4\|\Sigma\|}}(\operatorname{tr}\Sigma + \|\Sigma\|)2^{d/2+1} + \left(e^{-\frac{\widetilde{R}^2}{4\|\Sigma\|}}\sqrt{\operatorname{tr}\Sigma}2^{d/2+1}\right)^2 \\
&\le e^{-\frac{\widetilde{R}^2}{4\|\Sigma\|}}(\operatorname{tr}\Sigma + \|\Sigma\|)2^{d+3} = \widetilde{E}(\Sigma,\widetilde{R})\,.
\end{aligned} \tag{F.1.17}$$

Taylor-Lagrange inequality now yields the full result since $\log f(0)/f(0) = 0$ and $\nabla \log f(0) = 0$. ∎

Lemma F.2. *Consider $X \sim \mathcal{N}(0,\sigma^2 I)$ a multivariate Gaussian distribution with $\sigma^2 > 0$. For $R > 0$, define the truncated Gaussian r.v. X_R by conditioning X to the ball $\overline{\mathbb{B}}(0,R)$. Define*

$$\mathcal{H}(p) \coloneqq \log\mathbb{E}\left[e^{\langle p, X_R\rangle}\right] \tag{F.1.18a}$$

$$\mathcal{L}(p) \coloneqq \mathcal{H}^*(p)\,, \tag{F.1.18b}$$

and

$$E(\sigma^2, R) \coloneqq e^{-\frac{R^2}{16\sigma^2}} 2^{d+3}(d+1) \tag{F.1.19}$$

and assume that $R > 0$ satisfies

$$R \geq 4\sigma\sqrt{(d+3)\log 2 + \log(d+1)}\,. \tag{F.1.20}$$

Then, for any $p \in \mathcal{X}$ such that $\|p\| \leq \frac{R}{2\sigma^2}$, $v \in \mathcal{X}$ such that $\|v\| \leq \frac{R}{4}$, it holds that

$$\big(1 - E(\sigma^2, R)\big)\frac{\sigma^2\|p\|^2}{2} \leq \bar{\mathcal{H}}(p) \leq \big(1 + E(\sigma^2, R)\big)\frac{\sigma^2\|p\|^2}{2} \tag{F.1.21a}$$

$$\big(1 - 2E(\sigma^2, R)\big)\frac{\|v\|^2}{2\sigma^2} \leq \bar{\mathcal{L}}(v) \leq \big(1 + 2E(\sigma^2, R)\big)\frac{\|v\|^2}{2\sigma^2}\,. \tag{F.1.21b}$$

Proof. First, let us show that, for any $r > 0$, $\nabla\mathcal{H}\big(\overline{\mathbb{B}}(0,r)\big)$ is an open ball centered at 0. Since the distribution of X_R is invariant by rotation, this set can be rewritten as

$$\nabla\mathcal{H}\big(\overline{\mathbb{B}}(0,r)\big) = \big\{v \in \mathcal{X} : \|v\| \in \big\{\|\nabla\mathcal{H}(p)\| : p \in \overline{\mathbb{B}}(0,r)\big\}\big\}\,. \tag{F.1.22}$$

But, by connectedness of $\overline{\mathbb{B}}(0,r)$ and continuity of $\nabla\mathcal{H}$, $\big\{\|\nabla\mathcal{H}(p)\| : p \in \overline{\mathbb{B}}(0,r)\big\}$ is an interval. Since $\nabla\mathcal{H}(0) = 0$, it is either $[0, \|\nabla\mathcal{H}(r)\|]$ or $[0, \|\nabla\mathcal{H}(-r)\|)$. $\nabla\mathcal{H}\big(\overline{\mathbb{B}}(0,r)\big)$ being compact and therefore closed, it must be the latter. Hence, we have shown that

$$\nabla\mathcal{H}\big(\overline{\mathbb{B}}(0,r)\big) = \overline{\mathbb{B}}\left(0, \sup_{\overline{\mathbb{B}}(0,r)} \|\nabla\mathcal{H}\|\right). \tag{F.1.23}$$

We apply Lemma F.1 with $\widetilde{R} \leftarrow \frac{R}{2}$. Note that our choice of R implies that $\widetilde{R}$ satisfies the condition of Lemma F.1 and, moreover, that $E(\sigma^2, R) \leq \frac{1}{2}$. Lemma F.1 directly implies the bound on $\bar{\mathcal{H}}$.

Consider $p \in \mathcal{X}$ such that $\|p\| \leq \frac{R}{2\sigma^2}$. Then, by Lemma F.1, we have that

$$\begin{aligned}\big\|\nabla\bar{\mathcal{H}}(p)\big\| &\geq \sigma^2\|p\| - e^{-\frac{R^2}{16\sigma^2}}2^{d+3}\big(\mathrm{tr}\big(\sigma^2 I\big) + \big\|\sigma^2 I\big\|\big)\|p\| \\ &= \sigma^2\|p\|(1 - E(\sigma^2, R)) \\ &\geq \frac{\sigma^2\|p\|}{2}\,,\end{aligned} \tag{F.1.24}$$

where we used that $E(\sigma^2, R) \leq \frac{1}{2}$. Therefore, we obtain that $\sup\big\{\big\|\nabla\bar{\mathcal{H}}(p)\big\| : p \in \overline{\mathbb{B}}\big(0, \frac{R}{2\sigma^2}\big)\big\} \geq \frac{R}{4}$ so that $\nabla\mathcal{H}\big(\overline{\mathbb{B}}\big(0, \frac{R}{2\sigma^2}\big)\big)$ contains $\overline{\mathbb{B}}(0, \frac{R}{4})$.

Take $p \in \overline{\mathbb{B}}(0, \frac{R}{4})$ which therefore belongs to $\nabla\mathcal{H}\big(\overline{\mathbb{B}}\big(0, \frac{R}{2\sigma^2}\big)\big)$. Therefore, $\bar{\mathcal{L}}(v)$ can be rewritten as

$$\bar{\mathcal{L}}(v) = \sup_{p \in \overline{\mathbb{B}}\left(0, \frac{R}{2\sigma^2}\right)} \langle v, p\rangle - \bar{\mathcal{H}}(p)\,. \tag{F.1.25}$$

Using Lemma F.1 again, we obtain that

$$\sup_{p \in \overline{\mathbb{B}}\left(0, \frac{R}{2\sigma^2}\right)} \langle v, p\rangle - \frac{\sigma^2}{2}(1 + E(\sigma^2, R))\|v\|^2 \leq \bar{\mathcal{L}}(v) \leq \sup_{p \in \overline{\mathbb{B}}\left(0, \frac{R}{2\sigma^2}\right)} \langle v, p\rangle - \frac{\sigma^2}{2}(1 - E(\sigma^2, R))\|v\|^2\,, \tag{F.1.26}$$

so that, since $E(\sigma^2, R) \leq \frac{1}{2}$, we obtain that

$$\frac{\|v\|^2}{2\sigma^2(1 + E(\sigma^2, R))} \leq \bar{\mathcal{L}}(v) \leq \frac{\|v\|^2}{2\sigma^2(1 - E(\sigma^2, R))}\,. \tag{F.1.27}$$

Since, for $x \in [0, 1/2]$, both $\frac{1}{1+x} \geq 1 - 2x$ and $\frac{1}{1-x} \leq 1 + 2x$ hold, we get

$$\frac{\|v\|^2}{2\sigma^2}\big(1 - 2E(\sigma^2, R)\big) \leq \bar{\mathcal{L}}(v) \leq \frac{\|v\|^2}{2\sigma^2}\big(1 + 2E(\sigma^2, R)\big) \tag{F.1.28}$$

which concludes the proof. ∎

We will require the following technical lemma.

Lemma F.3. *Consider $v, w \in \mathcal{X}$ such that $0 < \|w\| \leq \frac{\mu R}{2}$ for some $R, \mu > 0$. Define,*

$$f(u) = \sup_{p \in \mathcal{X} : \|p\| \leq R} \langle p, u \rangle - \frac{\mu}{2}\|p\|^2 \,, \tag{F.1.29}$$

then, with $\lambda = \frac{\|v\|}{\|w\|}$,

$$\lambda f\Big(\frac{v}{\lambda} + w\Big) \leq f(v + w) \,. \tag{F.1.30}$$

Proof. Define $p \coloneqq \frac{1}{\mu}\big(\frac{v}{\lambda} + w\big)$ which has norm at most R since $\|w\| \leq \frac{R}{2\mu}$. Then, we have that

$$\begin{aligned}
\lambda f\Big(\frac{v}{\lambda} + w\Big) - f(v + w) &\leq \lambda\Big(\Big\langle p, \frac{v}{\lambda} + w\Big\rangle - \frac{\mu}{2}\|p\|^2\Big) - \Big(\langle p, v + w\rangle - \frac{\mu}{2}\|p\|^2\Big) \\
&= (\lambda - 1)\Big(\langle p, w\rangle - \frac{\mu}{2}\|p\|^2\Big) \\
&= (\lambda - 1) \times \frac{1}{\mu}\left(\frac{\langle v, w\rangle}{\lambda} + \|w\|^2 - \frac{1}{2}\left(2\|w\|^2 + 2\frac{\langle v, w\rangle}{\lambda}\right)\right) \\
&= 0 \,,
\end{aligned} \tag{F.1.31}$$

and our proof is complete. ∎

References

[1] Alongi, J. M. and Nelson, G. S. *Recurrence and Topology*, volume 85 of *Graduate Studies in Mathematics*. American Mathematical Society, 2007.

[2] Antonakopoulos, K., Mertikopoulos, P., Piliouras, G., and Wang, X. AdaGrad avoids saddle points. In *ICML '22: Proceedings of the 39th International Conference on Machine Learning*, 2022.

[3] Benaïm, M. Dynamics of stochastic approximation algorithms. In Azéma, J., Émery, M., Ledoux, M., and Yor, M. (eds.), *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pp. 1–68. Springer Berlin Heidelberg, 1999.

[4] Benaïm, M. and Hirsch, M. W. Dynamics of Morse-Smale urn processes. *Ergodic Theory and Dynamical Systems*, 15(6):1005–1030, December 1995.

[5] Bertsekas, D. P. and Tsitsiklis, J. N. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[6] Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *COLT '20: Proceedings of the 33rd Annual Conference on Learning Theory*, 2020.

[7] Brown, L. D. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, volume 9 of *Lecture Notes-Monograph Series*. Institute of Mathematical Statistics, 1986.

[8] Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G., and LeCun, Y. The loss surfaces of multilayer networks. In *AISTATS '15: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

[9] Coste, M. An introduction to o-minimal geometry, 1999.

[10] Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *NIPS '14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014.

[11] de Acosta, A. D. *Large Deviations for Markov Chains*. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, UK, 2022.

[12] Dembo, A. and Zeitouni, O. *Large Deviations Techniques and Applications*. Springer-Verlag, Berlin, 1998.

[13] Dieuleveut, A., Durmus, A., and Bach, F. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348–1382, June 2020.

[14] Douc, R., Moulines, E., Priouret, P., and Soulier, P. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer, 2018.

[15] Dupuis, P. Large deviations analysis of some recursive algorithms with state dependent noise. *The Annals of Probability*, 16(4):1509–1536, October 1988.

[16] Dupuis, P. and Kushner, H. J. Stochastic approximations via large deviations: Asymptotic properties. *SIAM Journal on Control and Optimization*, 23(5):675–696, September 1985.

[17] Eberle, A. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3-4):851–886, December 2016.

[18] Erdogdu, M. A., Mackey, L. W., and Shamir, O. Global non-convex optimization with discretized diffusions. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

[19] Feng, Y., Gao, T., Li, L., Liu, J.-G., and Lu, Y. Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation. *Communications in Mathematical Sciences*, 18 (1):163–188, 2020.

[20] Freidlin, M. I. and Wentzell, A. D. *Random Perturbations of Dynamical Systems*. Springer, 1 edition, 1998.

[21] Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points – Online stochastic gradient for tensor decomposition. In *COLT '15: Proceedings of the 28th Annual Conference on Learning Theory*, 2015.

[22] Gulinsky, O. V. and Veretennikov, A. Y. *Large Deviations for Discrete-Time Processes with Averaging*. De Gruyter, 1993.

[23] Gürbüzbalaban, M., Şimşekli, U., and Zhu, L. The heavy-tail phenomenon in SGD. In *ICML '21: Proceedings of the 38th International Conference on Machine Learning*, 2021.

[24] Hernández-Lerma, O. and Lasserre, J.-B. *Markov Chains and Invariant Probabilities*. Birkhäuser, Basel, 2003.

[25] Hodgkinson, L. and Mahoney, M. Multiplicative noise and heavy tails in stochastic optimization. In *ICML '21: Proceedings of the 38th International Conference on Machine Learning*, 2021.

[26] Hsieh, Y.-P., Mertikopoulos, P., and Cevher, V. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *ICML '21: Proceedings of the 38th International Conference on Machine Learning*, 2021.

[27] Hsieh, Y.-P., Karimi, M. R., Krause, A., and Mertikopoulos, P. Riemannian stochastic optimization methods avoid strict saddle points. In *NeurIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

[28] Hu, W., Li, C. J., Li, L., and Liu, J.-G. On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*, 4(1):3–32, 2019.

[29] Ioffe, A. D., Tikhomirov, V. M., and Makowski, K. *Theory of Extremal Problems*. North Holland, Amsterdam, NL, 1979.

[30] Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in SGD. https://arxiv.org/abs/1711.04623, 2017.

[31] Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *ICML '17: Proceedings of the 34th International Conference on Machine Learning*, 2017.

[32] Kallenberg, O. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, 2021.

[33] Kawaguchi, K. Deep learning without poor local minima. In *NIPS '16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

[34] Kiefer, J. and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[35] Kifer, Y. *Random Perturbations of Dynamical Systems*. Birkhäuser, Boston, MA, 1988.

[36] Kifer, Y. A discrete-time version of the Wentzell-Freidlin theory. *Annals of Probability*, 18(4):1676–1692, October 1990.

[37] Lan, G. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

[38] Landau, L. D. and Lifshitz, E. M. Statistical physics. In *Course of Theoretical Physics*, volume 5. Pergamon Press, Oxford, 1976.

[39] Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176(1):311–337, February 2019.

[40] Li, L. and Wang, Y. A sharp uniform-in-time error estimate for stochastic gradient Langevin dynamics. https://arxiv.org/abs/2207.09304, 2022.

[41] Li, L. and Wang, Y. On uniform-in-time diffusion approximation for stochastic gradient descent. https://arxiv.org/abs/2207.04922, 2022.

[42] Li, Q., Tai, C., and E, W. Stochastic modified equations and adaptive stochastic gradient algorithms. In *ICML '17: Proceedings of the 34th International Conference on Machine Learning*, 2017.

[43] Li, Q., Tai, C., and E, W. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

[44] Li, Z., Lyu, K., and Arora, S. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[45] Li, Z., Wang, T., and Arora, S. What happens after SGD reaches zero loss? –A mathematical framework. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021.

[46] Li, Z., Wang, T., and Yu, D. Fast mixing of stochastic gradient descent with normalization and weight decay. In *NeurIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

[47] Liu, S., Papailiopoulos, D., and Achlioptas, D. Bad global minima exist and SGD can reach them. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[48] Ljung, L. Analysis of recursive stochastic algorithms. *IEEE Trans. Autom. Control*, 22(4):551–575, August 1977.

[49] Lu, Y., Balasubramanian, K., Volgushev, S., and Erdogdu, M. A. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. In *NeurIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[50] Lytras, I. and Mertikopoulos, P. Tamed Langevin sampling under weaker conditions. https://arxiv.org/abs/2405.17693, 2024.

[51] Majka, M. B., Mijatović, A., and Szpruch, Ł. Nonasymptotic bounds for sampling algorithms without log-concavity. *The Annals of Applied Probability*, 30(4):1534–1581, 2020.

[52] Mertikopoulos, P., Hallak, N., Kavis, A., and Cevher, V. On the almost sure convergence of stochastic gradient descent in non-convex problems. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[53] Mertikopoulos, P., Hsieh, Y.-P., and Cevher, V. A unified stochastic approximation framework for learning in games. *Mathematical Programming*, 203:559–609, January 2024.

[54] Mignacco, F. and Urbani, P. The effective noise of stochastic gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083405, 2022.

[55] Mignacco, F., Krzakala, F., Urbani, P., and Zdeborová, L. Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[56] Mori, T., Ziyin, L., Liu, K., and Ueda, M. Power-law escape rate of SGD. In *ICML '22: Proceedings of the 39th International Conference on Machine Learning*, 2022.

[57] Pavasovic, K. L., Durmus, A., and Şimşekli, U. Approximate heavy tails in offline (multi-pass) stochastic gradient descent. In *NeurIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

[58] Pemantle, R. Nonconvergence to unstable points in urn models and stochastic aproximations. *Annals of Probability*, 18(2):698–712, April 1990.

[59] Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *COLT '17: Proceedings of the 30th Annual Conference on Learning Theory*, 2017.

[60] Rigollet, P. and Hütter, J.-C. High-dimensional statistics. https://arxiv.org/abs/2310.19244, 2023.

[61] Robbins, H. and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics*, 22: 400–407, 1951.

[62] Tutte, W. T. *Graph theory*. Cambridge University Press, 2001.

[63] van den Dries, L. and Miller, C. Geometric categories and *o*-minimal structures. *Duke Mathematical Journal*, 84(2), August 1996.

[64] Veiga, R., Remizova, A., and Macris, N. Stochastic gradient flow dynamics of test risk and its exact solution for weak features. https://arxiv.org/abs/2402.07626, 2024.

[65] Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[66] Wang, Y. and Wang, Z. Three-stage evolution and fast equilibrium for SGD with non-degerate critical points. In *ICML '22: Proceedings of the 39th International Conference on Machine Learning*, 2022.

[67] Wojtowytsch, S. Stochastic gradient descent with noise of machine learning type. Part I: Discrete time analysis. *Journal of Nonlinear Science*, 33(3):45, 2023.

[68] Xie, Z., Sato, I., and Sugiyama, M. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021.

[69] Yang, J., Hu, W., and Li, C. J. On the fast convergence of random perturbations of the gradient flow. *Asymptotic Analysis*, 122(3-4):371–393, 2021.

[70] Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S. P., and Glynn, P. W. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1):687–716, 2020.

[71] Ziyin, L., Li, H., and Ueda, M. Law of balance and stationary distribution of stochastic gradient descent. https://arxiv.org/abs/2308.06671, 2023.