On the projected Aubry set of the rate function associated with large deviations for stochastic approximations

Henrik Hult, Adam Lindhe, Pierre Nyquist

August 18, 2023

Abstract

In this article, we look at the problem of minimizing an action potential that arises from large devation theory for stochastic approximations. The solutions to the minimizing problem satisfy, in the sense of a viscosity solution, a Hamilton-Jacobi equation. From weak KAM theory, we know that these viscosity solutions are characterised by the projected Aubry set. The main result of this paper is that, for a specific rate function corresponding to a stochastic approximation algorithm, we prove that the projected Aubry set is equal to the forward limit set to the limit ODE.

1 Introduction

The theory of large deviations provides a powerful set of techniques to study rare events of stochastic processes. In many instances the large deviations rate function associated with a sequence of stochastic processes $\{X^n(t); t \in [0,T]\}$ on the space of continuous functions can be written as an action functional of the form

$$I(\psi) = \int_0^T L(\psi(s), \dot{\psi}(s))ds, \tag{1.1}$$

where ψ is an absolutely continuous function from $[0,T]$ to $\mathbb{R}^d$ with derivative $\dot{\psi}$ and $L : \mathbb{R}^{2d} \to \mathbb{R}$ is the local rate function, such that $v \mapsto L(x,v)$ is convex for all $x \in \mathbb{R}^d$. Markov processes provide a rich source of examples of processes for which the rate function is of this form, see e.g. [2, 3, 8, 15].

The study of rare events leads naturally to variational problems of minimizing the action functional under appropriate boundary conditions. For example, consider the probability that the process X^n exits an open set $\Omega \subset \mathbb{R}^d$ before time T, conditioned on $X(t) = x$, for $0 \le t < T$, $x \in \Omega$; written $P_{t,x}(X^n(T) \notin \Omega)$. If the rate function associated with X^n is given by (1.1), then the large deviations rate of this probability is given by

$$U(t,x) = \inf_{\psi} \Big\{ \int_t^T L(\psi(s), \dot{\psi}(s))ds, \psi(t) = x, \psi(T) \notin \Omega \Big\},$$

where the infimum is taken over all absolutely continuous functions. Intuitively, for large n, the large deviations principle leads to the asymptotic approximation

$$P_{t,x}(X^n(T) \notin \Omega) \approx e^{-nU(t,x)}.$$

The value function, $U(t,x)$, provides a measure of the rarity of the event that X^n leaves Ω, whereas the minimizing trajectory gives information about the most likely path that leads to the rare event. Since U is the value function of a variational problem, it satisfies, in the sense of a viscosity solution, a Hamilton-Jacobi terminal value problem of the form

$$\begin{cases} U_t(t,x) - H(x, -DU(t,x)) = 0, & (t,x) \in [0,T) \times \Omega, \\ U(T,x) = 0, & x \in \partial\Omega, \end{cases} \tag{1.2}$$

where H is the Fenchel-Legendre transform of L, see e.g. [8], U_t is the time derivative and D is the gradient.

A variational problem similar to (1.2) defines the action functional

$$M(t,y;x) = \inf_{\psi}\Big\{ \int_0^t L(\psi(s), \dot{\psi}(s))ds, \psi(0) = x, \psi(t) = y\Big\}, \quad t > 0\ x, y \in \mathbb{R}^d,$$

where the infimum is taken over all absolutely continuous functions $\psi : [0,t] \to \mathbb{R}^d$. This action functional is a well-studied object in large deviations theory and control-theory and may be interpreted as a cost that measures the exponential rate of decay of the probability of transiting from x to a neighborhood of y in time t, see for example [10, 9] and the references therein. In the weak KAM and dynamical systems literature it is often referred to as Mather's action functional, see the overview paper [13] and references therein, even though this functional was known well before the papers by Mather.

Transitions costs over arbitrary time intervals may be studied by considering Peierl's barrier and the Mañé potential. For a given $c \in \mathbb{R}$, Peierl's barrier refers to the limit of the action functional given by

$$h^c(x,y) = \liminf_{t\to\infty}\{M(t,y;x) + ct\},$$

and the projected Aubry set as the points for which Peierl's barrier vanishes,

$$\mathcal{A}^c = \{x \in \mathbb{R}^d : h^c(x,x) = 0\}.$$

The Mañé potential at level $c \in \mathbb{R}$ given by the value of the variational problem

$$S^c(x,y) = \inf_{\psi,t}\Big\{ \int_0^t c + L(\psi(s), \dot{\psi}(s))ds, \psi(0) = x, \psi(t) = y\Big\}, \quad x, y \in \mathbb{R}^d,$$

where, again, the infimum is taken over all absolutely continuous functions $\psi : [0,\infty) \to \mathbb{R}^n$ and $t > 0$, see [12]. For given x and c, consider the function $y \mapsto S^c(x,y)$. When this function is continuous, it is a viscosity subsolution of the stationary Hamilton-Jacobi equation

$$H(y, DS(y)) = c, \quad y \in \mathbb{R}^d.$$

From the definition of the Mañé potential it is elementary to show that

$$S^c(x,y) = \inf_{t>0}\{M(t,y;x) + ct\}.$$

For coercive Hamiltonians, the Mañé potential is a viscosity solution to the stationary Hamilton-Jacobi equation on $\mathbb{R}^d$ and the Aubry set is important because all viscosity solutions to the stationary Hamilton-Jacobi equation depends essentially on the Mañé potential $y \mapsto S^c(x,y)$ where x is in the projected Aubry set, see Theorem 3.6 below.

In this paper, we study Peierl's barrier, the projected Aubry set and the Mañé potential related to the action functional given by the rate function associated with the large deviations principle for stochastic approximations. The stochastic approximation sequence $\{X_k\}_{n\in\mathbb{N}_0}$ is of the Robbins-Munro type, with state-dependent Markovian noise $\{Y_k\}_{k\in\mathbb{N}}$, starting from X_0 and satisfying the recursion,

$$X_{k+1} = X_k + \frac{1}{k+1} g(X_k, Y_{k+1}), \quad k \geq 0,$$

where $g : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_x}$ is the update function, $\{Y_n\}_{n\in\mathbb{N}_0}$ is the noise sequence starting from Y_0, and, for every $k \in \mathbb{N}_0$ and $A \in \mathcal{B}(\mathbb{R}^{d_y})$

$$P(Y_{k+1} \in A | X_k, Y_k) = \rho_{X_k}(Y_k, A)$$

with $\rho_x(y,\cdot) \in \mathcal{P}(\mathbb{R}^{d_y})$ for any $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$. Stochastic approximation algorithms are frequently used for training statistical and machine learning models via stochastic gradient descent and persistent contrastive divergence algorithms, in adaptive Markov chain Monte Carlo methods and in statistical physics computations such as the Wang-Landau algorithm.

The large deviations principle provides information about the asymptotic behaviour of the stochastic approximation $\{X_n\}_{n\in\mathbb{N}}$ for large values of n. More specifically, for each $n \in \mathbb{N}$ and $x_0 \in \mathbb{R}^d$, define a process $\{X_k^n\}_{k\geq n}$ that follows the same recursive iterations but starts from the n-th step recursion. To be precise, let $X_n^n = x_0$ and for $k \geq n$

$$X_{k+1}^n = X_k^n + \frac{1}{n} g(X_k^n, Y_{n+k+1}). \tag{1.3}$$

We consider a family of continuous interpolations of $\{X_k^n\}_{k\geq n}$: for each n, $X^n = \{X^n(t) : t \in [0,T]\}$ is given by $X^n(\frac{k}{n} - \frac{1}{n}) = X_{n+k}^n$ for $k = 0, 1, \dots$, and for intermediate time points t, $X^n(t)$ is defined by a piece-wise linear interpolation. Note that, for each n, $X^n \in C_{x_0}([0,T] : \mathbb{R}^{d_x})$. The Markov kernel, ρ_x, for the noise sequence is assumed to be ergodic with invariant measure π_x. We define the limit function $\bar{g}(x)$ as

$$\bar{g}(x) = E_{Y\sim\pi_x}[g(x,Y)].$$

From standard results in stochastic approximation, we expect that for large n the function $X^n(t)$ should behave similarly to the solution of the ordinary differential equation,

$$\dot{x}(t) = \bar{g}(x), \tag{1.4}$$

called the limit ODE. The large deviations principle provides information about the rate at which the stochastic approximation deviates significantly from the solution to the limit ODE.

Under certain assumptions on the update function g, the family of Markov kernels $\{\rho_x\}$ and the step size sequence $\{\epsilon_k\}$, the linear interpolation $\{X^n\}$ of the stochastic approximation satisfies a large deviations principle with rate function (1.1) where the local rate function, e.g. the Lagrangian, is of the form

$$L(x,\beta) \doteq \inf_{\mu} \left\{ \inf_{\gamma \in \mathcal{M}(\mu)} R(\gamma(dy \times dz) || \mu(dx) \otimes \rho_x(y,dz)) : \beta = \int g(x,z)\mu(dz) \right\}, \tag{1.5}$$

where μ is a probability distribution $\mathbb{R}^d$, $\mathcal{M}(\mu)$ are distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with μ as marginal distributions. By understanding the variational problems appearing in the large deviations rate function associated with stochastic approximation we gain insights into fundamental properties of convergence, including typical and abnormal behavior of stochastic approximations. Our main result shows that

The paper is organized as follows. In Section 2 the Mañé potential and stationary Hamilton-Jacobi equations are introduced and we establish some relevant properties for a general Lagrangian. In Section 3 a stochastic approximation algorithm and its corresponding rate function is introduced and we characterise the projected Aubryb set for the local rate function to this rate function.

2 The Mañé potential and the stationary Hamilton-Jacobi equation

We begin by introducing the Mañé potential and establish some of its properties, as well as its relation to the stationary Hamilton-Jacobi equation. Throughout the paper we make the following assumption: the *Langrangian* $L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a locally bounded measurable function that is convex in the second coordinate. The *Hamiltonian* H is the Fenchel-Legendre transform of L,

$$H(x,p) = \sup_{v} \{\langle p, v\rangle - L(x,v)\}, \tag{2.1}$$

and by convex duality it follows that

$$L(x,v) = \sup_{p} \{\langle p, v\rangle - H(x,p)\}.$$

2.1 The Mañé potential

Originally introduced by Mañé in [12], the *Mañé potential at level* $c \in \mathbb{R}$, is the function $S^c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by

$$S^c(x,y) = \inf_{\psi,t} \Big\{ \int_0^t c + L(\psi(s), \dot{\psi}(s))ds, \psi(0) = x, \psi(t) = y \Big\}, \quad x,y \in \mathbb{R}^d, \tag{2.2}$$

where the infimum is taken over all $t > 0$ and absolutely continuous paths $\psi : [0,\infty) \to \mathbb{R}^d$. Since L is locally bounded it follows that $S^c(x,y) < \infty$, for all $x,y \in \mathbb{R}^d$ and $c < \infty$. It is possible that S^c is identically $-\infty$ for small c. Indeed, if $L(x,v) = \frac{1}{2}|v|^2$ and $c < 0$, then it follows from the definition (2.2) that $S^c(x,y) = -\infty$ for all $x,y \in \mathbb{R}^d$, by taking a path ψ for x to y that remains at x for an arbitrary long time interval, and letting the interval length tend to infinity. For completeness, some elementary and well-known properties of S^c are established for the specific setting considered here.

Proposition 2.1 *Let S^c be given by (2.2). The following properties hold.*

(i) For each $x,y \in \mathbb{R}^d$, the function $c \mapsto S^c(x,y)$ is nondecreasing.

(ii) For each $c \in \mathbb{R}$, the function $(x,y) \mapsto S^c(x,y)$ satisfies the triangle inequality:

$$S^c(x,z) \le S^c(x,y) + S^c(y,z), \quad x,y,z \in \mathbb{R}^d. \tag{2.3}$$

(iii) If $S^c(x_0,y_0) = -\infty$ for some $x_0,y_0 \in \mathbb{R}^d$ and $c \in \mathbb{R}$, then $S^c(x,y) = -\infty$ for all $x,y \in \mathbb{R}^d$.

(iv) If $S^c > -\infty$, then $S^c(x,x) = 0$, for each $x \in \mathbb{R}^d$.

Throughout the paper c_L denotes the infimum over all c such that $S^c > -\infty$.

Proof. (i) follows immediately from the definition of the Mañé potential. (ii) For the triangle inequality, if $S^c(x,z) = -\infty$ there is nothing to prove. Suppose $S^c(x,z) > -\infty$. Then $S^c(x,y) > -\infty$ and $S^c(y,z) > -\infty$ as well, for otherwise, if $S^c(x,y) = -\infty$, then there exists, for each $N > 0$, a $t_N > 0$ and an absolutely continuous path ψ_N with $\psi_N(0) = x$ and $\psi_N(t_N) = y$ such that

$$S^c(x,y) \le \int_0^{t_N} c + L(\psi_N(s), \dot{\psi}_N(s))ds \le -N.$$

Let $\tau > 0$ and φ be any absolutely continuous path with $\varphi(0) = y$ and $\varphi(\tau) = z$ and $\int_0^\tau c + L(\varphi(s), \dot{\varphi}(s))ds =: C < \infty$. Then, by concatenating ψ_N and φ as

$$\psi_N(s)I\{0 \le s \le t_N\} + \varphi(s - t_N)I\{t_N < s \le t_N + \tau\}$$

it follows that

$$S^c(x,z) \le \int_0^{t_N} c + L(\psi_N(s), \dot{\psi}_N(s))ds + \int_0^\tau c + L(\varphi(s), \dot{\varphi}(s))ds \le -N + C.$$

By sending $N \to \infty$ it follows that $S^c(x,z) = -\infty$, which is a contradiction. Consequently, $S^c(x,y) > -\infty$. A similar argument shows that $S^c(y,z) > -\infty$.

To proceed with the proof of the triangle inequality, take an arbitrary $\epsilon > 0$, and select $t_1, t_2 > 0$ and absolutely continuous paths ψ_1, ψ_2 with $\psi_1(0) = x$, $\psi_1(t_1) = y$, $\psi_2(0) = y$ and $\psi_2(t_2) = z$ such that

$$S^c(x,y) \ge \int_0^{t_1} c + L(\psi_1(s), \dot{\psi}_1(s))ds - \frac{\epsilon}{2},$$

$$S^c(y,z) \ge \int_0^{t_2} c + L(\psi_2(s), \dot{\psi}_2(s))ds - \frac{\epsilon}{2}.$$

Concatenate the two trajectories by

$$\psi(s) = \psi_1(s) I\{0 \le s \le t_1\} + \psi_2(s - t_1) I\{t_1 < s \le t_1 + t_2\}.$$

It follows that

$$\begin{aligned} S^c(x,y) + S^c(y,z) &\ge \int_0^{t_1} c + L(\psi_1(s), \dot{\psi}_1(s)) ds \\ &\quad + \int_0^{t_2} c + L(\psi_2(s), \dot{\psi}_2(s)) ds - \epsilon \\ &= \int_0^{t_1+t_2} c + L(\psi(s), \dot{\psi}(s)) ds - \epsilon \\ &\ge S^c(x,z) - \epsilon. \end{aligned}$$

Since $\epsilon > 0$ is arbitrary the triangle inequality follows.

(iii) follows from the triangle inequality.

To prove (iv), take $x \in \mathbb{R}^d$ and let $\epsilon > 0$, $h > 0$ be such that $h(c + L(x,0)) < \epsilon$ and $\psi(s) = x$ for each $0 \le s \le h$. By definition of the Mañé potential,

$$S^c(x,x) \le h(c + L(x,0)) < \epsilon.$$

Since $\epsilon > 0$ is arbitrary it follows that $S^c(x,x) \le 0$. The reverse inequality, $S^c(x,x) \ge 0$, follows from the triangle inequality. ■

2.2 The stationary Hamilton-Jacobi equation

Given a Hamiltonian $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $c \in \mathbb{R}$, the stationary Hamilton-Jacobi equation is

$$H(y, DS(y)) = c, \quad y \in \mathbb{R}^d. \tag{2.4}$$

A continuous function $S : \mathbb{R}^d \to \mathbb{R}$ is a *viscosity subsolution (supersolution)* of the stationary Hamilton-Jacobi equation (2.4) if, for every function $v \in C^\infty(\mathbb{R}^d)$,

$$\left.\begin{array}{l} \text{if } S - v \text{ has a local maximum (minimum) at } y_0 \in \mathbb{R}^d, \\ \text{then } H(y_0, Dv(y_0)) \le c \quad (\ge c). \end{array}\right\} \tag{2.5}$$

Such a function S is a *viscosity solution* if it is both a viscosity subsolution and a viscosity supersolution.

The *Mañé critical value* is the infimum over c for which (2.4) admits a viscosity subsolution. With some abuse of notation it will be denoted by c_H. The critical value admits the lower bound

$$c_H \ge \sup_y \inf_p H(y,p). \tag{2.6}$$

Indeed, if (2.4) admits a viscosity subsolution U^c at level c, then for every y there is a $v \in C^\infty(\mathbb{R}^d)$ such that $U^c - v$ has a local maximum at y and $\inf_p H(y,p) \le H(y, Dv(y)) \le c$. The claim follows by taking supremum over y. Examples where $c_H = \sup_y \inf_p H(y,p)$ are provided below.

The Mañé potential (2.2) is well studied within weak KAM theory, where it is commonly assumed that the Hamiltonian is uniformly superlinear: for each $K \geq 0$ there exists $C(K) \in \mathbb{R}$ such that $H(y,p) \geq K|p| - C(K)$ for each y, p. Under such an assumption there exist critical viscosity subsolutions, that is, there exists a global viscosity subsolution to (2.4) for $c = c_H$, see [6, 5]. In this paper uniform superlinearity is not assumed. In contrast, it is assumed that the Hamiltonian is given by the Fenchel-Legendre transform of a Lagrangian L, as in (2.1), and consequently $p \mapsto H(y,p)$ is convex in p, for every $y \in \mathbb{R}^d$. For instance, the Hamiltonian associated with the unit rate Poisson process, which is of the form

$$H(p) = e^p - 1, \quad p \in \mathbb{R},$$

is covered by our assumptions. For this choice of H the Mañé critical value is $c_H = -1$, but there can be no critical subsolution S as it would have to satisfy $DS(y) = -\infty$ almost everywhere, see Example 2.3 below.

The following properties of the Mañé potential are well known and similar statements appear in [5, 6, 7], see also the lecture notes [4, 1]. However, our assumptions on the Hamiltonian are different and thus a proof is included for completeness.

Proposition 2.2 *Assume* (2.1)*. Take $c \in \mathbb{R}$, $x \in \mathbb{R}^d$ and suppose that the function $y \mapsto S^c(x,y)$ is continuous. The following statements hold.*

(i) Suppose that $S^c > -\infty$. Then $y \mapsto S^c(x,y)$ is a viscosity subsolution to $H(y, DS(y)) = c$ on $\mathbb{R}^d$ and a viscosity solution on $\mathbb{R}^d \setminus \{x\}$.

(ii) For each $y \in \mathbb{R}^d$, $S^c(x,y) = \sup_{S \in \mathcal{S}_x^c} S(y)$, where $\mathcal{S}_x^c$ is the collection of all continuous viscosity subsolutions to $H(y, DS(y)) = c$ that vanish at x.

Recall that c_L is the infimum over c such that $S^c > -\infty$. Take $x \in \mathbb{R}^d$ and suppose that, for each $c > c_L$, the function $y \mapsto S^c(x,y)$ is continuous. For $c > c_H$ there exist viscosity subsolutions to (2.4) and by Proposition 2.2(ii) it follows that $S^c > -\infty$. Consequently, $c_H \geq c_L$. Similarly, for $c < c_H$ there are no subsolutions and by Proposition 2.2(i) $S^c = -\infty$, which implies $c_H \leq c_L$. This proves the following.

Corollary 2.3 *Take $x \in \mathbb{R}^d$ and suppose that, for each $c > c_L$, the function $y \mapsto S^c(x,y)$ is continuous. Then $c_H = c_L$.*

Before proceeding to the proof of Proposition 2.2 we state an important lemma that can be interpreted as a dynamic programming property of the Mañé potential.

Lemma 2.4 *Suppose that $S^c > -\infty$. For any $x, y_0 \in \mathbb{R}^d$ with $y_0 \neq x$ and $\epsilon > 0$ there exist $0 < \delta < |x - y_0|$, y with $|y - y_0| < \delta$, $h > 0$ and an absolutely continuous path ψ with $\psi(0) = y$, $\psi(h) = y_0$, and $|\psi(s) - y_0| < \delta$ for all $s \in [0,h]$, such that*

$$S^c(x, y_0) \geq S^c(x,y) + \int_0^h \Big(c + L(\psi(s), \dot{\psi}(s))\Big)\, ds - \epsilon.$$

Proof. Given $x, y_0 \in \mathbb{R}^d$ with $x \neq y_0$ and $\epsilon > 0$, take $t > 0$ and an absolutely continuous function φ with $\varphi(0) = x$, $\varphi(t) = y_0$ such that

$$S^c(x, y_0) \geq \int_0^t (c + L(\varphi(s), \dot{\varphi}(s)))\, ds - \epsilon.$$

Let $0 < \delta < |x - y_0|$ and take $h > 0$ such that $|\varphi(s) - y_0| < \delta$ for each $s \in [t-h, t]$. With $y = \varphi(t-h)$ and $\psi(s) = \varphi(s + t - h)$, $s \in [0, h]$, it follows that

$$\begin{aligned} S^c(x, y_0) &\geq \int_0^t (c + L(\varphi(s), \dot{\varphi}(s)))\, ds - \epsilon \\ &= \int_0^{t-h} (c + L(\varphi(s), \dot{\varphi}(s)))\, ds + \int_{t-h}^t (c + L(\varphi(s), \dot{\varphi}(s)))\, ds - \epsilon \\ &\geq S^c(x, y) + \int_0^h \Big(c + L(\psi(s), \dot{\psi}(s))\Big)\, ds - \epsilon. \end{aligned}$$

This completes the proof. ■

Proof of Proposition 2.2. Proof of (i). Suppose that $S^c > -\infty$, take $x \in \mathbb{R}^d$ and suppose that $y \mapsto S^c(x, y)$ is continuous. First we prove the viscosity subsolution property. For $v \in C^\infty(\mathbb{R}^d)$, suppose that $S^c(x, \cdot) - v$ has a local maximum at y_0 and, contrary to what we want to show, that $H(y, Dv(y)) - c \geq \theta > 0$ for $|y - y_0| \leq \delta$, for some $\delta > 0$. We may assume that δ is sufficiently small that

$$S^c(x, y) - v(y) \leq S^c(x, y_0) - v(y_0), \quad \text{for } |y - y_0| \leq \delta.$$

Take any y with $|y - y_0| \leq \delta$ and consider any absolutely continuous path ψ such that $\psi(0) = y$, $\psi(h) = y_0$ and $|\psi(s) - y_0| \leq \delta$ for all $s \in [0, h]$. By the triangle inequality (2.3) and the last inequality

$$\begin{aligned} 0 &\geq S^c(x, y_0) - S^c(x, y) - \int_0^h \Big(c + L(\psi(s), \dot{\psi}(s))\Big)\, ds \\ &\geq v(y_0) - v(y) - \int_0^h \Big(c + L(\psi(s), \dot{\psi}(s))\Big)\, ds \\ &= \int_0^h \left(\frac{d}{ds} v(\psi(s)) - L(\psi(s), \dot{\psi}(s)) - c\right) ds \\ &= \int_0^h \Big(\langle Dv(\psi(s)), \dot{\psi}(s)\rangle - L(\psi(s), \dot{\psi}(s)) - c\Big)\, ds. \end{aligned}$$

We may assume that $\dot{\psi}$ is chosen such that, using the conjugacy between H and L,

$$H(\psi(s), Dv(\psi(s))) \leq \langle Dv(\psi(s)), \dot{\psi}(s)\rangle - L(\psi(s), \dot{\psi}(s)) + \frac{\theta}{2},$$

for all $s \in [0, h]$. Then

$$\frac{\theta h}{2} \geq \int_0^h (H(\psi(s), Dv(\psi(s))) - c)\, ds \geq \theta h,$$

which is a contradiction. Thus, it must hold that $H(y_0, Dv(y_0)) \leq c$.

Next, we prove the supersolution property on $\mathbb{R}^d \setminus \{x\}$. Take $v \in C^\infty(\mathbb{R}^d)$ and suppose $S^c(x, \cdot) - v$ has a local minimum at $y_0 \neq x$ and, contrary to what we want to show, that $H(y, Dv(y)) - c \leq -\theta < 0$ for $|y - y_0| \leq \delta$, for some $\delta > 0$. We may assume that δ is sufficiently small that $|x - y_0| > \delta$ and

$$S^c(x, y) - v(y) \geq S^c(x, y_0) - v(y_0), \quad \text{for } |y - y_0| \leq \delta.$$

By Lemma 2.4 we may select y with $|y - y_0| \leq \delta$ and an absolutely continuous path ψ such that $\psi(0) = y$, $\psi(h) = y_0$ and $|\psi(s) - y_0| \leq \delta$ for all $s \in [0, h]$, with the property that

$$S^c(x, y_0) \geq S^c(x, y) + \int_0^h c + L(\psi(s), \dot{\psi}(s)) ds - \frac{\theta h}{2}.$$

The last inequality implies that

$$\begin{aligned}
\frac{\theta h}{2} &\geq S^c(x, y) - S^c(x, y_0) + \int_0^h \Big(c + L(\psi(s), \dot{\psi}(s)) \Big) \, ds \\
&\geq v(y) - v(y_0) + \int_0^h \Big(c + L(\psi(s), \dot{\psi}(s)) \Big) \, ds \\
&= \int_0^h \Big(-\frac{d}{ds} v(\psi(s)) + L(\psi(s), \dot{\psi}(s)) + c \Big) \, ds \\
&= \int_0^h \Big(-\langle Dv(\psi(s)), \dot{\psi}(s) \rangle + L(\psi(s), \dot{\psi}(s)) + c \Big) \, ds \\
&\geq \int_0^h - \Big(H(\psi(s), Dv(\psi(s))) - c \Big) ds.
\end{aligned}$$

We conclude that

$$-\frac{\theta h}{2} \leq \int_0^h \left(H(\psi(s), Dv(\psi(s))) - c \right) ds \leq -\theta h.$$

This is a contradiction and thus it must indeed hold that $H(y_0, Dv(y_0)) \geq c$, which completes the proof of (i).

Proof of (ii). Let $c \in \mathbb{R}$. If there are no viscosity subsolutions at level c, then by (i) $S^c = -\infty$ and $\mathcal{S}_x^c = \emptyset$, which implies that $\sup_{S \in \mathcal{S}_x^c} S(y) = -\infty$ as well. If there exist continuous viscosity subsolutions at level c, take $x \in \mathbb{R}^d$ and let S be a continuous viscosity subsolution of $H(y, DS(y)) = c$ on $\mathbb{R}^d$. It is sufficient to show that for any $y \in \mathbb{R}^d$, $t > 0$ and absolutely continuous function ψ with $\psi(0) = x$ and $\psi(t) = y$,

$$S(y) - S(x) \leq \int_0^t \Big(c + L(\psi(s), \dot{\psi}(s)) \Big) \, ds. \tag{2.7}$$

To show (2.7), fix $t > 0$, $y \in \mathbb{R}^d$, an absolutely continuous path ψ with $\psi(0) = x$ and $\psi(t) = y$ and take an arbitrary $\epsilon > 0$. For every $s \in [0, t]$, let $v_s \in C^\infty(\mathbb{R}^d)$ be such that $S - v_s$ has a local maximum at $\psi(s)$. Then, there exists $\delta_s > 0$ such that

$$S(z) - v_s(z) \leq S(\psi(s)) - v_s(\psi(s)), \qquad \text{for } |z - \psi(s)| < \delta_s,$$

and consequently that

$$S(z) - S(\psi(s)) \leq v_s(z) - v_s(\psi(s)), \qquad \text{for } |z - \psi(s)| < \delta_s. \tag{2.8}$$

By continuity of H and Dv_s we may, in addition, assume that δ_s is sufficiently small that

$$H(z, Dv_s(z)) \leq c + \frac{\epsilon}{t}, \qquad \text{for } |z - \psi(s)| < \delta_s.$$

For every $s \in [0,t]$, let $h_s > 0$ be such that $|\psi(u) - \psi(s)| < \delta_s$ for every u with $|u - s| < h_s$. This is possible due to the continuity of ψ. The union

$$[0, h_0) \cup \bigcup_{s \in (0,t]} (s, s + h_s),$$

is an open cover of $[0,t]$. Since $[0,t]$ is compact there is a finite subcover, which we can assume is of the form

$$[0, h_0) \cup \bigcup_{k=1}^{n-1} (s_k, s_k + h_{s_k}),$$

where $0 = s_0 < s_1 < \cdots < s_{n-1} < s_n = t$. Since the finite union is a subcover, it must hold that $s_{k-1} < s_k < s_{k-1} + h_{s_{k-1}}$ for each $k = 1, \ldots, n$. It follows that, using (2.8) and the conjugacy between H and L,

$$\begin{aligned}
S(y) - S(x) &= \sum_{k=1}^{n} S(\psi(s_k)) - S(\psi(s_{k-1})) \\
&\leq \sum_{k=1}^{n} v_{s_{k-1}}(\psi(s_k)) - v_{s_{k-1}}(\psi(s_{k-1})) \\
&= \sum_{k=1}^{n} \int_{s_{k-1}}^{s_k} \langle Dv_{s_{k-1}}(\psi(s)), \dot{\psi}(s) \rangle ds \\
&\leq \sum_{k=1}^{n} \int_{s_{k-1}}^{s_k} \Big(H(\psi(s), Dv_{s_{k-1}}(\psi(s))) + L(\psi(s), \dot{\psi}(s)) \Big) ds \\
&\leq \sum_{k=1}^{n} \int_{s_{k-1}}^{s_k} \Big(c + \frac{\epsilon}{t} + L(\psi(s), \dot{\psi}(s)) \Big) ds \\
&= \epsilon + \int_0^t \Big(c + L(\psi(s), \dot{\psi}(s)) \Big) ds.
\end{aligned}$$

Since $\epsilon > 0$ was arbitrary the claim follows. ∎

We proceed by computing Mañé's critical value, c_H for some Hamiltonians arising in the theory of large deviations of stochastic processes; in all three examples there is equality in the lower bound for c_H.

Example 2.1 (Critical diffusion process) *Let $U : \mathbb{R}^d \to \mathbb{R}$ be a potential function and $b(y) = -DU(y)$. Consider the Hamiltonian $H(y,p) = \langle b(y), p\rangle + \frac{1}{2}|p|^2$. Then $c_H = \sup_y \inf_p H(y,p) = -\frac{1}{2}\inf_y |b(y)|^2$. Indeed, from (2.6), $c_H \geq -\frac{1}{2}\inf_y |b(y)|^2$ and U is a subsolution to $H(y, DS(y)) = -\frac{1}{2}\inf_y |b(y)|^2$, which implies $c_H \leq -\frac{1}{2}\inf_y |b(y)|^2$. In particular, if $DU(y) = 0$ for some y, then $c_H = 0$. In this setting the Mañé potential can be viewed as a generalization of Freidlin and Wentzell's quasi-potential, described in [10, Ch. 4].*

Example 2.2 (Birth-and-death process) *Consider an interval $(a,b) \subset \mathbb{R}$ and functions $\mu : (a,b) \to [0,\infty)$, $\lambda : (a,b) \to [0,\infty)$ satisfying $\int_a^b \log(\sqrt{\mu(y)/\lambda(y)})dy < \infty$. Consider the Hamiltonian*

$$H(y,p) = \lambda(y)(e^p - 1) + \mu(y)(e^{-p} - 1).$$

In this case $c_H = \sup_y \inf_p H(y,p) = -\inf_y(\sqrt{\mu(y)} - \sqrt{\lambda(y)})^2$. To see this, recall from (2.6) that $c_H \geq -\inf_y(\sqrt{\mu(y)} - \sqrt{\lambda(y)})^2$. A subsolution of

$$H(y, DS(y)) = -\inf_y(\sqrt{\mu(y)} - \sqrt{\lambda(y)})^2,$$

is given by

$$U(y) = \int_a^y \log(\sqrt{\mu(z)/\lambda(z)})dz.$$

Indeed,

$$H(y, DU(y)) = -(\sqrt{\mu(y)} - \sqrt{\lambda(y)})^2 \leq -\inf_y(\sqrt{\mu(y)} - \sqrt{\lambda(y)})^2.$$

Example 2.3 (Pure birth process) *Let $\lambda : [0,\infty)^d \to [0,\infty)^d$ and put*

$$H(y,p) = \sum_{j=1}^d \lambda_j(y)(e^{p_j} - 1).$$

In this case $c_H = \sup_y \inf_p H(y,p) = -\inf_y \sum_{j=1}^d \lambda_j(y) =: -\lambda_*$. Indeed, from (2.6) it follows that $c_H \geq -\lambda_*$ and for any $c \in (-\lambda_*, 0)$ and $\alpha \leq \log(1 + c/\lambda_*)$, the function $\alpha\langle 1, y\rangle$ is a subsolution to $H(y, DS(y)) = c$, which implies $c_H \leq -\lambda_*$.

We end this subsection by proving a sufficient condition for the continuity of $y \mapsto S^c(x,y)$.

Proposition 2.5 *Suppose that the Lagrangian L is continuous at $(y,0)$ for each $y \in \mathbb{R}^d$. Then, for each $x \in \mathbb{R}^d$ and $c > c_L$ the function $y \mapsto S^c(x,y)$ is continuous.*

Proof. Take $y_0 \in \mathbb{R}^d$, $c > c_L$ and $\epsilon > 0$. To prove continuity at y_0 we show that there exists a $\delta > 0$ such that $|y - y_0| < \delta$ implies

$$S^c(x, y_0) \leq S^c(x,y) + \epsilon, \tag{2.9}$$

$$S^c(x,y) \leq S^c(x,y_0) + \epsilon. \tag{2.10}$$

We begin to prove (2.9). By assumption L is continuous at $(y_0, 0)$ and we may select δ' such that $L(y_0+z, v) \leq L(y_0, 0)+1$ for all $|z| < \delta'$ and $|v| < \delta'$. Pick $h > 0$ such that $h(c+L(y_0, 0)+1) < \epsilon/2$ and let $\delta = h\delta'$. For y such that $|y - y_0| < \delta$, take $t > h$ and an absolutely continuous path ψ with $\psi(0) = x$, $\psi(t-h) = y$ such that

$$S^c(x,y) \geq \int_0^{t-h} \Big(c + L(\psi(s), \dot{\psi}(s))\Big)\, ds - \frac{\epsilon}{2},$$

and $\dot{\psi}(s) = h^{-1}(y_0 - y)$ for $t - h \leq s \leq t$. Then,

$$\begin{aligned} S^c(x, y_0) &\leq \int_0^{t-h} \Big(c + L(\psi(s), \dot{\psi}(s))\Big)\, ds + \int_{t-h}^{t} \Big(c + L(\psi(s), \dot{\psi}(s))\Big)\, ds \\ &\leq S^c(x,y) + \frac{\epsilon}{2} + h(c + L(y_0, 0) + 1) \\ &\leq S^c(x,y) + \epsilon, \end{aligned}$$

by the choice of h. The proof of (2.10) is similar. ■

3 Large deviations for stochastic approximations and the projected Aubry set

In this section, we will study the Lagrangian appearing as the local rate function in a large deviations principle of stochastic approximations and characterise the projected Aubry set.

3.1 The Lagrangian associated with stochastic approximations

We are interested in the Lagrangian on the form

$$L(x,\beta) \doteq \inf_{\mu} \left\{ \inf_{\gamma \in \mathcal{M}(\mu)} R(\gamma(dy \times dz) || \mu(dx) \otimes \rho_x(y, dz)) : \beta = \int g(x,z)\mu(dz) \right\}, \tag{3.1}$$

where μ is a probability distribution $\mathbb{R}^{d_y}$, $\mathcal{M}(\mu)$ are distributions on $\mathbb{R}^{d_y} \times \mathbb{R}^{d_y}$ with μ as marginal distributions, $R(\cdot\|\cdot)$ is the relative entropy between probability distributions, $g(x,z)$ is the stochastic approximation update function from $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ to $\mathbb{R}^{d_x}$ and $\rho_x(y, dz)$ is a Markov kernel that depends on x with the following assumptions:

Assumption 3.1

(I) For every $\alpha \in \mathbb{R}^{d_x}$,

$$\sup_{x \in \mathbb{R}^{d_x}} \sup_{y \in \mathbb{R}^{d_y}} \left(\log \int_{\mathbb{R}^{d_y}} e^{\langle \alpha, g(x,z) \rangle} \rho_x(y, dz) \right) < \infty.$$

(ii) $L(x,\beta)$ is continuous in (x,β).

In [11] it is proven that under additional assumptions on g and $\{\rho_x\}$, the Lagrangian (3.1) is the local rate function of the stochastic approximation (1.3). However, for the main results of this paper we will not need such additional assumptions. The continuity of L will imply that its convex dual, H, is coercive. See [14].

The Lagrangian (3.1) has a number of useful properties, summarized in the following lemma.

Lemma 3.2 *Assume the Lagrangian $L(x,\beta)$ is on the form given by equation (3.1) satisfies assumption 3.1 then the following holds*

(i) For all x, $L(x,\beta)$ is convex in β.

(ii) $L(x,\beta)$ is super-linear. i.e for all $x \in \mathbb{R}^{d_x}$ and all $K > 0$ there exists a constant $C(K)$, that depends only on K, such that

$$L(x,\beta) \geq K\|\beta\| + C(K),$$

for all $\beta \in \mathbb{R}^{d_x}$.

Proof. Property (i) and (ii) are proved in [11]. Here we prove (iii). By Assumption 3.1, for all $K > 0$ there is a constant $\tilde{C}(K)$ such that,

$$\sup_{x\in\mathbb{R}^{d_x}} \sup_{y\in\mathbb{R}^{d_y}} \left(\int_{\mathbb{R}^{d_y}} e^{K\|g(x,z)\|} \rho_x(y,dz) \right) < \tilde{C}(K).$$

For a given $\epsilon > 0$, $\beta \in \mathbb{R}^{d_x}$ and $x \in \mathbb{R}^{d_x}$, choose measures $\mu(dy)$ and $q(y,dz)$ such that μ is invariant measure to q,

$$L(x,\beta) + \epsilon > R(q \otimes \mu || \rho_x \otimes \mu),$$

and

$$\beta = \int_{\mathbb{R}^{d_y}} g(x,y)\mu(dy).$$

Now for any β we have

$$\begin{aligned}\|\beta\| &= \left\| \int_{\mathbb{R}^{d_y}} g(x,y)\mu(dy) \right\| \leq \int_{\mathbb{R}^{d_y}} \|g(x,y)\| \, \mu(dy) = \int_{\mathbb{R}^{d_y}} \int_{\mathbb{R}^{d_y}} \|g(x,y)\| \, q(z,dy)\mu(dz) \\ &= \int_{\mathbb{R}^{d_y}} \int_{\mathbb{R}^{d_y}} \|g(x,y)\| \, \frac{q(z,dy)\mu(dz)}{\rho_x(z,dy)\mu(dz)} \rho_x(z,dy)\mu(dz).\end{aligned}$$

Now using the inequality $ab \leq e^{Ka} + \frac{1}{K}(b\log(b) + b - 1)$, with $a = \|g(x,y)\|$ and $b = \frac{q(z,dy)\mu(dz)}{\rho_x(z,dy)\mu(dz)}\rho_x(z,dy)$ yields

$$\begin{aligned}\|\beta\| &\leq \int_{\mathbb{R}^{d_y}} \int_{\mathbb{R}^{d_y}} e^{K\|g(x,y)\|} \rho_x(z,dy)\mu(dz) + \frac{1}{K} R(q \otimes \mu || \rho_x \otimes \mu) \\ &\leq \sup_{x\in\mathbb{R}^{d_x}} \sup_{y\in\mathbb{R}^{d_y}} \left(\int_{\mathbb{R}^{d_y}} e^{K\|g(x,z)\|} \rho_x(y,dz) \right) + \frac{1}{K}(L(x,\beta) + \epsilon) \\ &\leq \tilde{C}(K) + \frac{1}{K}(L(x,\beta) + \epsilon).\end{aligned}$$

With $C(K) = -K\tilde{C}(K) - \epsilon$ we have that

$$L(x, \beta) \geq K\|\beta\| + C(K),$$

which proves the superlinearity of $L(x, \beta)$. ■

The superlinearity is an important property because it yields compactness properties for minimizing curves to the Lagrangian.

Theorem 3.3 *Let L be a Lagrangian that is convex and superlinear, and let K be a compact set. Then the following subset of absolutely continuous paths*

$$\left\{\psi \in \mathcal{AC}([a,b]) : \psi([a,b]) \cap K \neq \emptyset, \int_a^b L(\psi(t), \dot{\psi}(t))dt \leq \gamma\right\},$$

is compact in the topology of uniform convergence for all $\gamma > 0$.

For proof see [5]. Now we proceed to establish the critical values for the Mañé potential associated with the Lagrangian (3.1).

Lemma 3.4 *For the Lagrangian specified in equation* (3.1) *and satisfying the assumptions 3.1, we have $c_H = c_L = 0$.*

Proof.

The equality between c_L and c_H follows from Corollary 2.3 and Proposition 2.5. Since the relative entropy is always larger than zero we have that $L(x, \beta) \geq 0$. This implies that $c_L \leq 0$. Let x be a point such that $L(x, 0) = 0$, these are the stationary points to the ODE 1.4. If $c_L < 0$ then the value $S^c(x, x)$ can be made arbitrarily small. This implies that $c_L \geq 0$ and therefore we have that $c_L = 0$. From now on we deal with the case of $c = 0$. ■

Since the Lagrangian $L \geq 0$, then $S^0(x, y) > -\infty$ and by Preposition 2.2 we have that $S^0(x, y)$ is a viscosity subsolution on $\mathbb{R}^{d_x}$ and a viscosity solution on $\mathbb{R}^{d_x}\backslash\{x\}$.

3.2 Characterisation of the projected Aubry set

There is a close connection between the projected Aubry set and viscosity solutions to the stationary Hamilton-Jacobi equations.

Lemma 3.5 *Let H be a coercive Hamiltonian and $c \geq c_H$. The following are equivalent:*

- *$x \in \mathcal{A}^c$.*
- *$y \to S^c(x, y)$ is a viscosity solution to $H(y, DS(y)) = c$ on $\mathbb{R}^d$.*

We know from Proposition 2.2 that, for each $x \in \mathbb{R}^{d_x}$, $y \mapsto S^c(x, y)$ is a viscosity solution on $\mathbb{R}^d\backslash\{x\}$, the projected Aubry set is precisely the points where this property extends to the whole space. The projected Aubry set is also important in characterising the viscosity solutions, for example if the underlying space is a compact closed manifold M, we have the following representation formula for viscosity solutions.

Theorem 3.6 *Given a coercive Hamiltonian H on a compact connected manifold M , all viscosity solutions $u : \mathbb{R}^d \to \mathbb{R}$ of the stationary Hamilton-Jacobi equation, $H(u, Du) = c_H$, satisfies*

$$u(x) = \inf_{y \in \mathcal{A}^c} [u(y) + S^{c_H}(y, x)].$$

See [5] for proof. Our contribution in this part is characterising the Aubry set for Lagrangian defined in (3.1). The forward set, at the point $x \in \mathbb{R}^d$, to the ODE (1.4) is

$$F(x) = \cap_{t>0} \overline{\{y(s) : s > t\}},$$

where $y(s)$ is a solution to the ODE with initial value $y(0) = x$. Define the total forward set F to be the union of all forward sets

$$F = \cup_{x \in \mathbb{R}^d} F(x).$$

The main theorem of this section connects the total forward set with the projected Aubry set.

Theorem 3.7 *The projected Aubry set $\mathcal{A}^0$ to the Lagrangian defined by* (3.1) *and satisfying the assumptions 3.1 is equal to the total forward set F.*

Proof. First, we prove the following, that for a trajectory satisfies $\psi(t)$, $\int_0^T L(\psi(t), \dot{\psi}(t))dt = 0$ is equivalent to that ψ is a solution almost everywhere to the ODE (1.4). From the definition of $L(x, \beta)$

$$L(x, \beta) = \inf_{\mu} \inf_{q : \mu q = q} \left\{ \int R(q(y, \cdot)||\rho_x(y, \cdot))\mu(dy) : \int g(x, y)\mu(dy) = \beta \right\},$$

we have that if $L(x, \beta) = 0$ iff $\beta = \bar{g}(x)$. if $\beta = \bar{g}(x)$ then we can take as the minimizing measures $\mu = \pi$ and $q = \rho_x$ which implies that $L(x, \beta) = 0$. If $L(x, \beta) = 0$ this means that for a minimizing measures μ and q we have that $q(y, \cdot) = \rho_x(y, \cdot)$ except for a μ-null set. This means that

$$\int \rho_x(y, dz)\mu(dy) = \int q(y, dz)\mu(dy) = \mu(dz),$$

which implies that μ is an invariant measure to ρ_x. Since the invariant measure is unique we have that $\mu = \pi$ and therefore that $\beta = E_\pi[g(x, Y)] = \bar{g}(x)$. So a trajectory ψ have zero cost iff $\dot{\psi}(x) = \bar{g}(x)$. Now we can prove that $x \in F \Rightarrow x \in \mathcal{A}^0$. If $x \in F$ then there is a point x_0 such that a solution to the ODE (1.4) $\psi(t)$ with the properties $\psi(0) = x_0$ and either $\psi(t) = x$ infinitely often or $\lim_{t\to\infty} \psi(t) = x$. In the case that $\psi(t)$ visits x infinitely often there is a increasing sequence of times t_k such that $\lim_{k\to\infty} t_k \to \infty$ and such that $\psi(t_k) = x$. Then

$$M(t_k - t_1, x, x) \leq \int_0^{t_k - t_1} L(\psi(t_1 + t), \dot{\psi}(t_1 + t))dt = 0,$$

which implies that $h(x, x) = 0$. In the case that $\psi(t)$ converges to x then since $\bar{g}(x)$ is uniformly continuous then $\dot{\psi}(t) \to 0$ and we have that x is a stationary point $\bar{g}(x) = 0$. Stationary points are obviously in the projected Aubry set.

Now we will assume that $x \in \mathcal{A}^0$. Then there exists trajectories ψ_k, increasing times t_k and numbers ϵ_k such that $\lim_k t_k \to \infty$, $\lim \epsilon_k = 0$ and

$$\int_0^{t_k} L(\psi_k(s), \dot{\psi}_k(s))ds < \epsilon_k.$$

By Theorem 3.3 the sequence $\{\psi_k\}$ is compact for every time interval. We can therefore construct a limit functionψ by

$$\psi(t) = \lim_{k\to\infty} \psi_k(t),$$

where we take the limit over a convergent subsequence. Now the function ψ satisfies

$$\int_0^T L(\psi(t), \dot{\psi}(t))dt = 0.$$

Therefore ψ must be a solution almost everywhere to the ODE 1.4 with initial value $\psi(0) = x$. Due to the uniform convergence we can for all $\epsilon > 0$ find a N such that for $k > N$, $\sup_{t\in[0,t_k]} \|\psi(t) - \psi_k(t)\| < \epsilon$. Since $\psi_k(t_k) = x$ we have that for every $\epsilon > 0$ we can find infinite times t_k such that $\|\psi(t_k) - x\| < \epsilon$. This implies that

$$x \in \cap_{t>0}\{\psi(s), s > t\} = F(x).$$

So x is in the total forward set F.

To further specify the projected Aubry set we take some insights from Lyaponov theory. If the limit function is on the form $\bar{g}(x) = -f_x(x)$ for some real valued function f, bounded from below and $f(x) \to \infty$ when $\|x\| \to \infty$, then the Forward limit set is all the stationary points $F = \mathcal{A}^0 = \{x : \bar{g}(x) = 0\}$. The conditions on $G(x)$ would for example be satisfied in a linear regression case with squared loss or logistic regression with cross-entropy loss. We conclude with some stochastic approximation algorithms where the projected Aubry set can be explicitly calculated.

Example 3.1 (Stochastic gradient descent) *Given some data $y = \{y_i\}_{i=1}^N$ and parameters x the task is to minimize a function $G(x) = \frac{1}{N}\sum_{i=1}^N G(y_i, x)$ with stochastic gradient descent (SGD). Assume that $G(y, x)$ is bounded from below and $G(x) \to \infty$ if $\|x\| \to \infty$. Given a estimate of the parameters x_n the next point is given by the procedure: sample I_{n+1} uniformly from $\{1, \dots, N\}$ and update x_{n+1} by*

$$x_{n+1} = x_n - \frac{1}{n+1}\nabla_x G(y_{I_{n+1}}, x).$$

In this setting we have that $g(x, I) = -\nabla_x G(y_I, x)$ and that the noise distribution ρ is the uniform measure over $\{1, \dots, N\}$. The Lagrangian is then given by

$$L(x, \beta) = \inf_\mu \left\{ \sum_{i=1}^N \log(N\mu_i)\mu_i : \beta = -\sum_{i=1}^N \nabla_x G(y_i, x)\mu_i \right\},$$

where the minimizing measure μ can be parameterized by a N-long probability vector. We also require $L(x, \beta) < \infty$ in a neighbourhood of $\beta = 0$. The corresponding Hamiltonian is given by

$$H(x, \alpha) = \log\left(\frac{1}{N}\sum_{i=1}^N e^{-\langle\alpha, \nabla_x G(y_i, x)\rangle}\right).$$

Since the limit function $\bar{g}(x) = -\nabla_x G(x)$ we have that the projected Aubry set is given by

$$\mathcal{A}^0 = \{x : \nabla_x G(x) = 0\},$$

all stationary points to the minimization problem.

Example 3.2 (Persistent contrastive divergence) *Consider a model with visible variables v, hidden variables h, model parameters x and probability density on the form*

$$p(v, h|x) = e^{-E(v,h;x)+F(x)},$$

where E is the energy and F is the free energy or normalisation constant

$$F(x) - \log \int e^{-E(v,h;x)} dh dx.$$

We also define the density for the hidden variables given the visible and the parameters as

$$p(h|v, x) = e^{-E(h,v;x)+F_H(x,v)},$$

where

$$F_H(v, x) = -\log \int e^{-E(v,h;x)} dh.$$

The persistent contrastive divergence algorithm minimizes the negative log-likelihood

$$G(x) = \sum_{i=1}^{N} \log p(v^i|x),$$

by a recursive updating scheme. The updating scheme relies on $N + 1$ Markov chains. The first N chains will be Markov chains in $y^1 = h$ and will approximate $p(h|v_i, x)$. We denote these with $\rho_x^{i,1}(y^1|dz^1)$ and construct these such that they have invariant distribution $p(h|v_i, x)$. The final Markov chain $\rho^2(y^2, dz^2)$ have variable $y^2 = (h, v)$ and have stationary distribution $p(v, h|x)$. Now given a parameter estimate X_n the next parameter is given by

$$X_{n+1} = X_n - \varepsilon_{n+1} \left(\nabla_x E(v^I, Y^1_{n+1}; x) - \nabla_x E(Y^2_{n+1}; x) \right).$$

The random index I is drawn uniformly from $\{1, \ldots, N\}$, Y^1_{n+1} is drawn from $\rho^{I,1}_{X_n}(Y^1_n, dz)$ and Y^2_{n+1} is drawn from $\rho_{X_n}(Y^2_n, dz)$. By identifying

$$g(x, y) = \nabla_x E(v^i, y^1; x) - \nabla_x E(y^2; x)$$

and the noise distribution as

$$\rho_x(y, dz)\rho_I(i) = \frac{1}{N} \sum_{j=1}^{N} \rho_x^{i,1}(y^1, dz)\rho_x(y^2, dz)\delta_{i=j},$$

we have a stochastic approximation update on the form of equation (1.3). The limit function is given by

$$\bar{g}(x) = \frac{1}{N} \sum_{i=1}^{N} \int \nabla_x E(v_i, h; x)p(h|v^i, x)dh - \int \nabla_x E(v, h; x)p(v, h; x)dvdh,$$

which is equal to the gradient of the log-likelihood, $\bar{g}(x) = L(x)$. Furthermore if the energy E is bounded from below and $E(h, v; x) \to \infty$ as $\|x\| \to \infty$ the projected Aubry set is given by

$$\mathcal{A}^0 = \{x : \nabla G(x) = 0\},$$

the minimizing points to the log-likelihood.

■

References

[1] P. Bernard. The lax-oleinik semi-group: a hamiltonian point of view. *CANPDE crash-course*, February:http://arxiv.org/pdf/1203.3569.pdf, 2011.

[2] A. Budhiraja and P. Dupuis. *Analysis and Approximation of Rare Events: Representations and Weak Convergence Methods.* Number 94 in Probability Theory and Stochastic Modelling. Springer-Verlag, New York, 2019.

[3] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications.* Stochastic Modelling and Applied Probability. Springer, New York, NY, second edition, 1998.

[4] A. Fathi. Weak kam from a pde point of view: viscosity solutions of the hamilton-jacobi equation and the aubry set. *Lecture notes from Course CANPDE*, 17-19 February:https://www.ceremade.dauphine.fr/ pbernard/enseignement/m2/fathi.pdf, 2011.

[5] A. Fathi. *Weak KAM Theorem in Lagrangian Dynamics.* Cambridge Studies in Advanced Mathematics. Cambrdige University Press, 2014.

[6] A. Fathi and Maderna E. Weak kam theorem on non compact manifolds. *Nonlinear Differential Equations and Applications NoDEA*, 14:1–27, 2007.

[7] A. Fathi and A. Siconolfi. Existence of c^1 critical subsolutions of the hamilton-jacobi equation. *Invent. Math.*, 155:363–388, 2004.

[8] J. Feng and T. G. Kurtz. *Large deviations for stochastic processes.* Mathematical surveys and monographs. American Mathematical Society, Providence, RI, first edition, 2006.

[9] W. H. Fleming and H.M. Soner. *Controlled Markov processes and viscosity solutions.* 1992.

[10] M. I. Freidlin and A. D. Wentzell. *Random perturbations of dynamical systems.* Springer, New York, NY, 1984.

[11] H. Hult, A. Lindhe, P. Nyquist, and G. Wu. A weak convergence approach to large deviations for stochastic approximations. *preprint*, 2023.

[12] R. Mañé. Langrangian flows: the dynamics of globally minimizing orbits. *Bull. Brazilian Mathematical Society*, 28(2):141–153, 1997.

[13] J. N. Mather. Variational construction of connecting orbits. *Ann. Inst. Fourier*, 43:1349–1386, 1993.

[14] A.W. Roberts and D.E. Varberg. *Convex Functions.* 1976.

[15] A. Shwartz and A. Weiss. *Large deviations for performance analysis.* Stochastic Modeling Series. Chapman & Hall, London, 1995. Queues, communications, and computing, With an appendix by Robert J. Vanderbei.