HAL
open science

# Stochastic mirror descent dynamics and their convergence in monotone variational inequalities

Panayotis Mertikopoulos, Mathias Staudigl

## To cite this version:

Panayotis Mertikopoulos, Mathias Staudigl. Stochastic mirror descent dynamics and their convergence in monotone variational inequalities. Journal of Optimization Theory and Applications, 2018, 179 (3), pp.838-867. 10.1007/s10957-018-1346-x . hal-01643343

**HAL Id: hal-01643343**

**https://hal.science/hal-01643343v1**

Submitted on 9 Oct 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# STOCHASTIC MIRROR DESCENT DYNAMICS AND THEIR CONVERGENCE IN MONOTONE VARIATIONAL INEQUALITIES

PANAYOTIS MERTIKOPOULOS* AND MATHIAS STAUDIGL♯

Abstract. We examine a class of stochastic mirror descent dynamics in the context of monotone variational inequalities (including Nash equilibrium and saddle-point problems). The dynamics under study are formulated as a stochastic differential equation driven by a (single-valued) monotone operator and perturbed by a Brownian motion. The system's controllable parameters are two variable weight sequences that respectively pre- and post-multiply the driver of the process. By carefully tuning these parameters, we obtain global convergence in the ergodic sense, and we estimate the average rate of convergence of the process. We also establish a large deviations principle showing that individual trajectories exhibit exponential concentration around this average.

## 1. Introduction

Discrete and continuous dynamical systems governed by maximal monotone operators play an important role in optimization, game theory, equilibrium, fixed-point theory, partial differential equations, among many others. The study of the relationship between continuous time and discrete time models is an active area of research, see [1] for a recent overview, and [2] for connections to accelerated methods. Viewing an iterative algorithm as a discrete version of a continuous dynamical system shed new light on the properties of this algorithm, offer Lyapunov functions which are useful for the asymptotic analysis, and suggest new classes of algorithms. A classical situation arises in the study of (projected) gradient descent dynamics and its connection with Cauchy's steepest descent algorithm, and the relation between Mirror descent algorithms and dynamical systems derived from Bregman projections [3–5]. This paper is concerned with the continuous time analysis of a random dynamical system which can be formally interpreted as the continuous-time version of a dual averaging algorithm developed by Nesterov [6, 7]. This algorithm performs simple gradient descent steps in the unconstrained dual space, and then projects the iterates from the dual space to the feasible set of the optimization

* Univ. Grenoble Alpes, CNRS, Inria, LIG, F-38000, Grenoble, France.

♯ Maastricht University, Department of Quantitative Economics, P.O. Box 616, NL–6200 MD Maastricht, The Netherlands.

*E-mail addresses*: panayotis.mertikopoulos@imag.fr, m.staudigl@maastrichtuniversity.nl.

2010 *Mathematics Subject Classification.* Primary 90C25, 60H10 and ; secondary 90C33, 90C47.

*Key words and phrases.* mirror descent, variational inequalities, saddle-point problems, stochastic differential equations.

P. Mertikopoulos was partially supported by the French National Research Agency (ANR) grant ORACLESS (ANR–16–CE33–0004–01) and the COST Action CA16228 "European Network for Game Theory" (GAMENET). The research of M. Staudigl is partially supported by the COST Action CA16228 "European Network for Game Theory" (GAMENET). .

problem. This projection is performed as in the classical mirror descent framework [3, 8]. In the realm of convex optimization, the dual step effectively computes a weighted average of the realized gradient of the objective function. This averaging steps makes this algorithm particularly suited for problems where only noisy information is available to the decision maker. For this reason dual averaging has been successively employed in machine learning and engineering [9, 10].

In the framework of convex optimization we have studied the resulting dynamical system in detail in [11]. This paper extends this dynamical approach considerably to the setting of general monotone variational inequality problems [12]. In this paper we study a time-continuous random dynamical system generated by a primal-dual method in mathematical programming originating from [7] based on monotone single-valued operators. The method combines a stochastic gradient step in the unconstrained dual space with a projection step in the primal. The main results of this paper are as follows: Firstly, we address the existence and uniqueness of continuous processes satisfying the defining equations of our method. Secondly, we provide a detailed analysis of the convergence of the trajectories in the deterministic as well as in the stochastic case. We provide a set of results on the convergence of the ergodic average of the solution trajectories, as well as convergence results of individual trajectories in case where the driving operator is strictly monotone.

Notation. Throughout this paper, $\mathcal{X}$ will denote a compact convex subset of an $n$-dimensional real space $\mathcal{V} \cong \mathbb{R}^n$ with norm $\|\cdot\|$. We will also write $\mathcal{Y} \equiv \mathcal{V}^*$ for the dual of $\mathcal{V}$, $\langle y, x\rangle$ for the canonical pairing between $y \in \mathcal{V}^*$ and $x \in \mathcal{V}$, and $\|y\|_* \equiv \sup\{\langle y, x\rangle : \|x\| \leq 1\}$ for the dual norm of $y$ in $\mathcal{V}^*$. We denote the relative interior of $\mathcal{X}$ by $\operatorname{ri}(\mathcal{X})$, and its boundary by $\operatorname{bd}(\mathcal{X})$. Finally, for an extended-real-valued function $f\colon \mathcal{V} \to \mathbb{R} \cup \{\infty\}$, we define its *domain* as $\operatorname{dom} f \equiv \{x \in \mathcal{V} : f(x) < \infty\}$.

## 2. Setup and preliminaries

2.1. **Variational inequalities.** Let $v\colon \mathcal{X} \to \mathcal{Y}$ be a Lipschitz continuous monotone map, i.e.

$$\begin{aligned} &\|v(x) - v(x')\|_* \leq L\|x - x'\|, \\ &\langle v(x) - v(x'), x - x'\rangle \geq 0, \end{aligned} \tag{H1}$$

for some $L > 0$ and for all $x, x' \in \mathcal{X}$. Throughout this paper, we will be interested in solving the *variational inequality* (VI) problem:

$$\text{Find } x_* \in \mathcal{X} \text{ such that } \langle v(x), x - x_*\rangle \geq 0 \text{ for all } x \in \mathcal{X}. \tag{VI}$$

Since $v$ is assumed continuous and monotone, this *weak* (or *Minty-type*) VI problem is equivalent to the *strong* (or *Stampacchia-type*) VI problem [13, 14]:

$$\text{Find } x_* \in \mathcal{X} \text{ such that } \langle v(x_*), x - x_*\rangle \geq 0 \text{ for all } x \in \mathcal{X}. \tag{VI$'$}$$

When we need to keep track of $\mathcal{X}$ and $v$ explicitly, we will refer to (VI) and/or (VI′) as $\operatorname{VI}(\mathcal{X}, v)$. The solution set of $\operatorname{VI}(\mathcal{X}, v)$ will be denoted as $\mathcal{X}_*$; by standard results, $\mathcal{X}_*$ is convex, compact and nonempty [12].

Below, we present a selected sample of examples and applications of variational inequality problems; for a more extensive discussion, see [12, 15, 16].

*Example* 2.1 (Convex optimization). Consider the problem

$$\begin{aligned} \text{minimize} \quad & f(x), \\ \text{subject to} \quad & x \in \mathcal{X}, \end{aligned} \tag{Opt}$$

where $f\colon \mathcal{X} \to \mathbb{R}$ is convex and continuously differentiable on $\mathcal{X}$. If $x_*$ is a solution of (Opt), first-order optimality gives

$$\langle \nabla f(x_*), x - x_* \rangle \geq 0 \quad \text{for all } x \in \mathcal{X}. \tag{2.1}$$

Since $f$ is convex, $v = \nabla f$ is monotone, so (Opt) is equivalent to $\mathrm{VI}(\mathcal{X}, \nabla f)$ [17].

*Example* 2.2 (Saddle-point problems). Let $\mathcal{X}^1 \subseteq \mathbb{R}^{n_1}$ and $\mathcal{X}^2 \subseteq \mathbb{R}^{n_2}$ be compact and convex, and let $U\colon \mathcal{X}^1 \times \mathcal{X}^2 \to \mathbb{R}$ be a smooth convex-concave function (i.e. $U(x^1, x^2)$ is convex in $x^1$ and concave in $x^2$). Then, the associated *saddle-point* (or *min-max*) problem is to determine the *value* of $U$, defined here as

$$\mathrm{val} = \min_{x^1 \in \mathcal{X}^1} \max_{x^2 \in \mathcal{X}^2} U(x^1, x^2) = \max_{x^2 \in \mathcal{X}^2} \min_{x^1 \in \mathcal{X}^1} U(x^1, x^2). \tag{Val}$$

That the value of $U$ is well-defined follows from von Neumann's minimax theorem. Moreover, letting

$$v(x^1, x^2) = \big(\nabla_{x^1} U(x^1, x^2), -\nabla_{x^2} U(x^1, x^2)\big), \tag{2.2}$$

it is easy to check that $v$ is monotone as a map from $\mathcal{X} \equiv \mathcal{X}^1 \times \mathcal{X}^2$ to $\mathbb{R}^{n_1+n_2}$ (because $U$ is convex in its first argument and concave in the second). Then, as in the case of (Opt), first-order optimality implies that the saddle-points of (Val) are precisely the solutions of the variational inequality $\mathrm{VI}(\mathcal{X}, v)$ [7].

*Example* 2.3 (Convex games). One of the main motivations for this paper comes from determining the Nash equilibria of games with convex cost functions. To state the problem, let $\mathcal{N} = \{1, \dots, N\}$ be a finite set of *players* and, for each $i \in \mathcal{N}$, let $\mathcal{X}^i \subseteq \mathbb{R}^{n_i}$ be a compact convex set of *actions* that can be taken by player $i$. Given an action profile $x = (x^1, \dots, x^N) \in \mathcal{X} \equiv \prod_i \mathcal{X}^i$, the cost for each player is determined by an associated *cost function* $c^i\colon \mathcal{X} \to \mathbb{R}$. The unilateral minimization of this cost leads to the notion of *Nash equilibrium* (NE), defined here as an action profile $x_* = (x_*^i)_{i \in \mathcal{N}}$ such that

$$c^i(x_*) \leq c^i(x^i; x_*^{-i}) \quad \text{for all } x^i \in \mathcal{X}^i,\ i \in \mathcal{N}. \tag{NE}$$

Of particular interest to us is the case where each $c^i$ is smooth and individually convex in $x^i$. In this case, the profile $v(x) = (v^i(x))_{i \in \mathcal{N}}$ of individual gradients $v^i(x) = \nabla_{x^i} c^i(x)$ forms a monotone map and, by first-order optimality, the Nash equilibrium problem (NE) boils down to solving $\mathrm{VI}(\mathcal{X}, v)$ [12, 18].

In the rest of this paper, we will consider two important special cases of (VI), namely:

(1) *Strictly monotone* problems, i.e. when

$$\langle v(x') - v(x), x' - x \rangle \geq 0 \qquad \text{with equality iff } x = x'. \tag{2.3}$$

(2) *Strongly monotone* problems, i.e. when

$$\langle v(x') - v(x), x' - x \rangle \geq \gamma \|x' - x\|^2 \quad \text{for some } \gamma > 0. \tag{2.4}$$

Clearly, strong monotonicity implies strict monotonicity (which in turns implies ordinary monotonicity). In the case of convex optimization problems, strict (respectively strong) monotonicity corresponds to strict (respectively strong) convexity of the problem's objective function. Under either refinement, (VI) admits a unique solution, which will be referred to as "the" solution of (VI).

2.2. **Stochastic mirror descent dynamics.** Mirror descent is an iterative optimization algorithm combining first-order oracle steps with a "mirror step" generated by a projection-type mapping.[1] The key ingredient defining this mirror step is a generalization of the Euclidean distance known as a "distance-generating" function:

**Definition 2.1.** We say that $h\colon \mathcal{X} \to \mathbb{R}$ is a *distance-generating function* on $\mathcal{X}$ if

*a)* $h$ is *continuous*.

*b)* $h$ is *strongly convex*, i.e. there exists some $\alpha > 0$ such that

$$h(\lambda x + (1-\lambda)x') \leq \lambda h(x) + (1-\lambda)h(x') - \frac{\alpha}{2}\lambda(1-\lambda)\|x - x'\|^2, \tag{2.5}$$

for all $x, x' \in \mathcal{X}$ and all $\lambda \in [0,1]$.

Given a distance-generating function on $\mathcal{X}$, its convex conjugate is given by

$$h^*(y) = \max_{x\in\mathcal{X}}\{\langle y, x\rangle - h(x)\}, \quad y \in \mathcal{Y}, \tag{2.6}$$

and the induced *mirror map* is defined as

$$Q(y) = \operatorname*{argmax}_{x\in\mathcal{X}}\{\langle y, x\rangle - h(x)\}. \tag{2.7}$$

Thanks to the strong convexity of $h$, $Q(y)$ is well-defined and single-valued for all $y \in \mathcal{Y}$. In particular, as illustrated in the examples below, it plays a role similar to that of a projection mapping:

*Example* 2.4 (Euclidean distance). If $h(x) = \frac{1}{2}\|x\|_2^2$, the induced mirror map is the standard Euclidean projector

$$Q(y) = \operatorname*{argmax}_{x\in\mathcal{X}}\left\{\sum\nolimits_{j=1}^n y_j x_j - \frac{1}{2}\sum\nolimits_{j=1}^n x_j^2\right\} = \operatorname*{argmin}_{x\in\mathcal{X}}\|x - y\|_2^2. \tag{2.8}$$

*Example* 2.5 (Gibbs–Shannon entropy). If $\mathcal{X} = \{x \in \mathbb{R}_+^n : \sum_{j=1}^n x_j = 1\}$ is the unit simplex in $\mathbb{R}^n$, the (negative) Gibbs-Shannon entropy $h(x) = \sum_{j=1}^n x_j \log x_j$ gives rise to the so-called *logit choice* map

$$Q(y) = \frac{(\exp(y_j))_{j=1}^n}{\sum_{k=1}^n \exp(y_k)}. \tag{2.9}$$

*Example* 2.6 (Fermi–Dirac entropy). If $\mathcal{X} = [0,1]^n$ is the unit cube in $\mathbb{R}^n$, the (negative) Fermi-Dirac entropy $h(x) = \sum_{j=1}^n [x_j \log x_j + (1-x_j)\log x_j]$ induces the so-called *logistic map*

$$Q(y) = \left(\frac{\exp(y_j)}{1+\exp(y_j)}\right)_{j=1}^n \tag{2.10}$$

For future reference, some basic properties of mirror maps are collected below:

**Proposition 2.2.** *Let $h$ be a distance-generating function on $\mathcal{X}$. Then, the induced mirror map $Q\colon \mathcal{Y} \to \mathcal{X}$ satisfies the following properties:*

*a)* *$x = Q(y)$ if and only if $y \in \partial h(x)$; in particular, $\operatorname{im} Q = \operatorname{dom} \partial h \subseteq \mathcal{X}$.*

*b)* *$h^*$ is continuously differentiable on $\mathcal{Y}$ and $\nabla h^*(y) = Q(y)$ for all $y \in \mathcal{Y}$.*

*c)* *$Q(\cdot)$ is $(1/\alpha)$-Lipschitz continuous.*

[1] For the origins of the method, see [8]; the specific variant we consider here is due to and is commonly referred to as "dual averaging" [7] or "lazy mirror descent (MD)" [19].

The properties reported above are fairly standard in convex analysis; for a proof, see e.g. [20, p. 217], [21, Theorem 23.5] and [17, Theorem 12.60(b)]. Of particular importance is the minimizing argument identity $\nabla h^* = Q$ which provides a quick way of calculating $Q$ in "prox-friendly" geometries (such as the examples discussed above).

Now, as mentioned above, mirror descent exploits the flexibility provided by a (not necessarily Euclidean) mirror map by using it to generate first-order steps along $v$. For concreteness, we will focus on the so-called "dual averaging" variant of mirror descent [7], defined here via the recursion

$$\begin{aligned} y_{t+1} &= y_t - \lambda_t v(x_t), \\ x_{t+1} &= Q(\eta_{t+1} y_{t+1}), \end{aligned} \tag{2.11}$$

where:

1) $t = 0, 1, \dots$ denotes the stage of the process.
2) $y_t$ is an auxiliary dual variable that aggregates first-order steps along $v$.[2]
3) $\lambda_t$ is a variable step-size parameter that *pre-multiplies* the input at each stage.
4) $\eta_t$ is a variable weight parameter that *post-multiplies* the dual aggregate $y_t$.[3]

Thus, descending to continuous time, we obtain the *mirror descent dynamics*

$$\begin{aligned} dy(t) &= -\lambda(t)\, v(x(t))\, dt, \\ x(t) &= Q(\eta(t) y(t)), \end{aligned} \tag{MD}$$

with $\eta(t)$ and $\lambda(t)$ serving the same role as before (but are now defined over all $t \geq 0$). In particular, our standing assumption for the parameters $\lambda$ and $\eta$ of (MD) will be that

$$\eta(t) \text{ and } \lambda(t) \text{ are positive, } C^1\text{-smooth and nonincreasing.} \tag{H2}$$

At a heuristic level, the assumptions above guarantee that the dual process $y(t)$ does not grow too large too fast, so blow-ups in finite time are not possible. Together with the basic convergence properties of the dynamics (MD), this is discussed in more detail in Section 3 below.

Now, the primary case of interest in our paper is when the oracle information for $v(x)$ in (MD) is subject to noise, measurement errors and/or other stochastic disturbances. To account for such perturbations, we will instead focus on the *stochastic mirror descent dynamics*

$$\begin{aligned} dY(t) &= -\lambda(t)\, [v(X(t))\, dt + \, dM(t)], \\ X(t) &= Q(\eta(t) Y(t)), \end{aligned} \tag{SMD}$$

where $M(t)$ is a continuous martingale with respect to some underlying stochastic basis $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$.[4] In more detail, we assume for concreteness that the stochastic disturbance term $M(t)$ is an Itô process of the form

$$dM(t) = \sigma(X(t), t) \cdot \, dW(t), \tag{2.12}$$

where:

[2]The usual initialization is $y_0 = 0$, $x_0 = Q(0) = \operatorname{argmin}_{x\in\mathcal{X}} h(x)$, but other initializations are possible.

[3]The name "dual averaging" alludes to the choice $\lambda_t = 1$, $\eta_t = 1/t$: under this choice of parameters, $x_t$ is a mirror projection of the "dual average" $y_t = t^{-1} \sum_{s=0}^{t-1} v(x_s)$.

[4]We tacitly assume here that the filtration $(\mathcal{F}_t)_{t\geq 0}$ satisfies the usual conditions of right continuity and completeness, and carries a standard $d$-dimensional Wiener process $(W(t))_{t\geq 0}$.

1) $W(t)$ is a $d$-dimensional Wiener process adapted to $\mathcal{F}_t$.
2) $\sigma(x,t)$ is an $n \times d$ matrix capturing the *volatility* of the noise process.

Heuristically, the volatility matrix of $M(t)$ captures the intensity of the noise process and the possible correlations between its components. For instance, when $d = n$ and $\sigma$ is the identity matrix, $M(t)$ is just a standard Wiener process: in this case, the increments of the noise are independent and identically distributed (i.i.d.) and they are not correlated across different components. Otherwise, if $\sigma$ is not diagonal, $M(t)$ could exhibit nontrivial correlations and/or other dependencies across components.

In terms of regularity, we will be assuming throughout that $\sigma(x,t)$ is measurable, bounded, and Lipschitz continuous in $x$. Formally, we posit that

$$\begin{aligned} \sup_{x,t} \|\sigma(x,t)\| &< \infty, \\ \|\sigma(x',t) - \sigma(x,t)\| &\leq \ell \|x' - x\|, \end{aligned} \tag{H3}$$

where $\ell > 0$ is a positive constant and

$$\|\sigma\| = \sqrt{\operatorname{tr}[\sigma\sigma^\top]} = \sqrt{\sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{d} |\sigma_{ij}|^2} \tag{2.13}$$

denotes the Frobenius (matrix) norm of $\sigma$. In particular, (H3) implies that there exists a positive constant $\sigma_* \geq 0$ such that

$$\|\sigma(x,t)\|^2 \leq \sigma_*^2 \quad \text{for all } x \in \mathcal{X},\, t \geq 0. \tag{2.14}$$

In what follows, it will be convenient to measure the intensity of the noise affecting (SMD) via $\sigma_*$; of course, when $\sigma_* = 0$, we recover the noiseless, deterministic dynamics (MD).

## 3. Deterministic analysis

To establish a reference standard, we first focus on the deterministic regime of (MD), i.e. when $M(t) \equiv 0$ in (SMD). We begin with a basic well-posedness result:

**Proposition 3.1.** *Under conditions* (H1) *and* (H2)*, the dynamical system* (MD) *admits a unique global solution.*

*Proof.* Let $A(t,y) = -\lambda(t) v(Q(\eta(t)y))$ for all $t \in \mathbb{R}_+$, $y \in \mathcal{Y}$. Clearly, $A(t,y)$ is jointly continuous in $t$ and $y$. Moreover, by (H2), $\lambda(t)$ has bounded first derivative and $\eta(t)$ is nonincreasing, so both $\lambda(t)$ and $\eta(t)$ are Lipschitz continuous. Finally, by (H1), $v$ is $L$-Lipschitz continuous, implying in turn that

$$\|A(t,y_1) - A(t,y_2)\|_* \leq \frac{L\eta(t)\lambda(t)}{\alpha} \|y_1 - y_2\|_* \quad \text{for all } y_1, y_2 \in \mathcal{Y}, \tag{3.1}$$

where $\alpha$ is the strong convexity constant of $h$ and we used Proposition 2.2 to estimate the Lipschitz constant of $Q$.

This shows that $A(t,y)$ is Lipschitz in $y$ for all $t$, so existence and uniqueness of local solutions follows from the Picard–Lindelöf theorem. Eq. (H2) further guarantees that the Lipschitz constant of $A(t,\cdot)$ can be chosen uniformly in $t$, so these solutions can be extended for all $t \geq 0$. ∎

Given existence and uniqueness of a unique global solution of the dual process, we can define a semi-flow $\phi(t,y) : [0,\infty) \times \mathcal{Y} \to \mathcal{Y}$ satisfying $\phi(0,y) = y$ and $\frac{\partial \phi(t,y)}{\partial t} = A(t, \phi(t,y))$ for all $(t,y) \in [0,\infty) \times \mathcal{Y}$. This semi-flow induces a Lipschitz continuous trajectory $\xi(t,y) = Q(\eta(t)\phi(t,y))$ on the primal space $\mathcal{X}$.

Now, to analyze the convergence of (MD), we will consider two "gap functions" quantifying the distance between $x(t)$ and the solution set of (VI):

- In the general case, we will focus on the "*dual gap*" (or "*merit*") function [12, 22]:
$$g(x) = \max_{x' \in \mathcal{X}} \langle v(x'), x - x' \rangle. \tag{3.2}$$
By (H1) and the compactness of $\mathcal{X}$, it follows that $g(x)$ is continuous, non-negative and convex; moreover, we have $g(x) = 0$ if and only if $x$ is a solution of $\mathrm{VI}(\mathcal{X}, v)$ [13, Proposition 3.1].
- For the saddle point problem described in Example 2.2, we instead look at the Nikaido-Isoda gap function [23]:
$$G(p^1, p^2) = \max_{x^2 \in \mathcal{X}^2} U(p^1, x^2) - \min_{x^1 \in \mathcal{X}^1} U(x^1, p^2). \tag{3.3}$$

Since $U$ is convex-concave, it is immediate that $G(p^1, p^2) \geq g(p^1, p^2)$, where the operator involved in the definition of the dual gap function is given by the saddle-point operator (2.2). However, it is still true that $G(p^1, p^2) = 0$ if and only if the pair $(p^1, p^2)$ is a saddle-point.[5] Since both gap functions vanish only at solutions of (VI), we will prove trajectory convergence by monitoring the decrease of the relevant gap over time. This is achieved by introducing the so-called *Fenchel coupling* [18], an auxiliary energy function defined here as

$$F(x, y) = h(x) + h^*(y) - \langle y, x \rangle \quad \text{for all } x \in \mathcal{X},\ y \in \mathcal{Y}, \tag{3.4}$$

where $h^*$ denotes the convex conjugate of $h$.

*Remark* 3.1. In a certain sense, the Fenchel coupling can be seen as a primal-dual extension of the well-known *Bregman divergence* [24, 25]:

$$D(x, z) = h(x) - h(z) - h'(z; x - z) \quad \text{for all } x, z \in \mathcal{X}. \tag{3.5}$$

More precisely, we have $F(x, y) \geq D(x, Q(y))$ with equality if and only if $Q(y)$ is interior [18, Prop. 4.3].

Some further key properties of $F$ are summarized in the following proposition (also proved in [18]):

**Proposition 3.2.** *Let $h$ be a distance-generating function on $\mathcal{X}$. Then:*

*a) $F(x, y) \geq 0$ with equality if and only if $x = Q(y)$; in particular, $F(x, y) \geq \frac{\alpha}{2}\|Q(y) - x\|^2$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$.*

*b) Viewed as a function of $y$, $F(x, y)$ is convex, differentiable, and its gradient is given by*
$$\nabla_y F(x, y) = Q(y) - x. \tag{3.6}$$

*c) For al $x \in \mathcal{X}$ and all $y, y' \in \mathcal{Y}$, we have*
$$F(x, y') \leq F(x, y) + \langle y' - y, Q(y) - x \rangle + \frac{1}{2\alpha}\|y' - y\|_*^2. \tag{3.7}$$

We are now in a position to state and prove our first convergence result for (MD). In the sequel, if there is no danger of confusion we will use the more concise notation $x(t) = \xi(t, y)$ and $y(t) = \phi(t, y)$ for the unique solution to (MD) with initial condition $y \in \mathcal{Y}$.

[5] Simply note that $\max_{x^2 \in \mathcal{X}^2} \min_{x^1 \in \mathcal{X}^1} U(x^1, x^2) \leq \min_{x^1 \in \mathcal{X}^1} \max_{x^2 \in \mathcal{X}^2} U(x^1, x^2)$, and $G(p^1, p^2) = 0$ implies that $\max_{x^2 \in \mathcal{X}^2} \min_{x^1 \in \mathcal{X}^1} U(x^1, x^2) \geq \min_{x^1 \in \mathcal{X}^1} \max_{x^2 \in \mathcal{X}^2} U(x^2, x^2)$.

Consider the averaged trajectory

$$\bar{x}(t) = \frac{\int_0^t \lambda(s)x(s)\,ds}{\int_0^t \lambda(s)\,ds} = \frac{1}{S(t)}\int_0^t \lambda(s)x(s)\,ds, \tag{3.8}$$

where we have set

$$S(t) = \int_0^t \lambda(s)\,ds. \tag{3.9}$$

We then have the following convergence guarantee:

**Proposition 3.3.** *Suppose that* (MD) *is initialized at* $y = 0$. *Then:*

$$g(\bar{x}(t)) \leq \frac{\mathcal{D}(h;\mathcal{X})}{\eta(t)S(t)}, \tag{3.10}$$

*where* $\bar{x}(t)$ *is the averaged trajectory constructed in* (3.8) *with* $x(t) = \xi(t,0)$, *and*

$$\mathcal{D}(h;\mathcal{X}) = \max_{x,x'\in\mathcal{X}}\{h(x') - h(x)\} = \max h - \min h. \tag{3.11}$$

*In particular, if* (VI) *is associated with a convex-concave saddle-point problem as in Example 2.2, we have the guarantee:*

$$G(\bar{x}(t)) \leq \frac{\mathcal{D}(h_1;\mathcal{X}^1) + \mathcal{D}(h_2;\mathcal{X}^2)}{\eta(t)S(t)}. \tag{3.12}$$

*In both cases,* $\bar{x}(t)$ *converges to the solution set of* $\mathrm{VI}(\mathcal{X},v)$ *whenever* $\lim_{t\to\infty}\eta(t)S(t) = \infty$.

*Proof.* Given some $p \in \mathcal{X}$, let

$$H_p(t) = \frac{1}{\eta(t)}F(p,\eta(t)y(t)) \tag{3.13}$$

denote the "$\eta$-deflated" Fenchel coupling between $p$ and $y(t) \equiv \phi(t,0)$. Then, by Proposition 3.2, a simple differentiation yields

$$H_p(t) - H_p(0) = -\int_0^t \lambda(s)\langle v(x(s)), x(s) - p\rangle ds - \int_0^t \frac{\dot{\eta}(s)}{\eta(s)^2}[h(p) - h(x(s))]ds, \tag{3.14}$$

and, after rearranging, we obtain

$$\begin{aligned}\int_0^t \lambda(s)\langle v(x(s)), x(s) - p\rangle\,ds &= H_p(0) - H_p(t) - \int_0^t \frac{\dot{\eta}(s)}{\eta(s)^2}[h(p) - h(x(s))]\,ds \\ &\leq H_p(0) + \mathcal{D}(h;\mathcal{X})\left(\frac{1}{\eta(t)} - \frac{1}{\eta(0)}\right).\end{aligned} \tag{3.15}$$

Now, let $x_c = \operatorname{argmin}\{h(x) : x \in \mathcal{X}\}$ denote the "prox-center" of $\mathcal{X}$. Since $\eta(0) > 0$ and $y(0) = 0$ by assumption, we readily get

$$H_p(0) = \frac{F(p,0)}{\eta(0)} = \frac{h(p) + h^*(0) - \langle 0, p\rangle}{\eta(0)} = \frac{h(p) - h(x_c)}{\eta(0)} \leq \frac{\mathcal{D}(h;\mathcal{X})}{\eta(0)}. \tag{3.16}$$

From the monotonicity of $v$, we further deduce that

$$g(\bar{x}(t)) \leq \frac{1}{S(t)}\max_{p\in\mathcal{X}}\int_0^t \lambda(s)\langle v(x(s)), x(s) - p\rangle\,ds. \tag{3.17}$$

Thus, substituting (3.16) in (3.15), maximizing over $p \in \mathcal{X}$ and plugging the result into (3.17) gives (3.10).

Suppose now that (VI) is associated to a convex-concave saddle-point problem as in Example 2.2. In this case. we can replicate the above analysis for each component $x^i(t)$, $i = 1, 2$, of $x(t)$ to obtain the basic bounds

$$\begin{aligned} \int_0^t \lambda(s)\langle \nabla_{x^1} U(x(s)), x^1(s) - p^1\rangle \, ds &\leq \frac{\mathcal{D}(h_1;\mathcal{X}^1)}{\eta(t)}, \\ \int_0^t \lambda(s)\langle -\nabla_{x^2} U(x(s)), x^2(s) - p^2\rangle \, ds &\leq \frac{\mathcal{D}(h_2;\mathcal{X}^2)}{\eta(t)}. \end{aligned} \tag{3.18}$$

Using the fact that $U$ is convex-concave, this leads to the value-based bounds

$$\begin{aligned} \int_0^t \lambda(s)[U(x(s)) - U(p^1, x^2(s))] \, ds &\leq \frac{\mathcal{D}(h_1;\mathcal{X}^1)}{\eta(t)}, \\ \int_0^t \lambda(s)[U(x^1(s), p^2) - U(x(s))] \, ds &\leq \frac{\mathcal{D}(h_2;\mathcal{X}^2)}{\eta(t)}. \end{aligned} \tag{3.19}$$

Summing these inequalities, dividing by $S(t)$, and using Jensen's inequality gives

$$U(\bar{x}^1(t), p^2) - U(p^1, \bar{x}^2(t)) \leq \frac{\mathcal{D}(h_1;\mathcal{X}^1) + \mathcal{D}(h_2;\mathcal{X}^2)}{\eta(t)S(t)}$$

The bound (3.12) then follows by taking the supremum over $p^1$ and $p^2$, and using the definition of the Nikaido–Isoda gap function. ■

The gap-based analysis of Proposition 3.3 can be refined further in the case of *strongly* monotone VI problems. In this case, we have:

**Proposition 3.4.** *Let $x_*$ denote the (necessarily unique) solution of a $\gamma$-strongly monotone problem* $\mathrm{VI}(\mathcal{X}, v)$. *Then,*

$$\|\bar{x}(t) - x_*\|^2 \leq \frac{\mathcal{D}(h;\mathcal{X})}{\gamma} \frac{1}{\eta(t)S(t)}. \tag{3.20}$$

*In particular, $\bar{x}(t)$ converges to $x_*$ whenever $\lim_{t\to\infty} \eta(t)S(t) = \infty$.*

*Proof.* By Jensen's inequality, the strong monotonicity of $v$ and the assumption that $x_*$ solves $\mathrm{VI}(\mathcal{X}, v)$, we have:

$$\begin{aligned} \gamma\|\bar{x}(t) - x_*\|^2 &\leq \frac{\gamma}{S(t)} \int_0^t \lambda(s)\|x(s) - x_*\|^2 \, ds && \text{(Jensen)} \\ &\leq \frac{1}{S(t)} \int_0^t \lambda(s)\langle v(x(s)) - v(x_*), x(s) - x_*\rangle \, ds && (\gamma\text{-monotonicity}) \\ &\leq \frac{1}{S(t)} \int_0^t \lambda(s)\langle v(x(s)), x(s) - x_*\rangle \, ds && \text{(optimality of } x_*) \\ &\leq \frac{\mathcal{D}(h;\mathcal{X})}{\eta(t)S(t)}, && (3.21) \end{aligned}$$

where the last inequality follows as in the proof of Proposition 3.3. The bound (3.20) is then obtained by dividing both sides by $\gamma$. ■

The two results above are in the spirit of classical ergodic convergence results for monotone VI problems as in [7, 26, 27]. In particular, taking $\eta(t) = \sqrt{L/(2\alpha)}$ and $\lambda(t) = 1/(2\sqrt{t})$ gives the upper bound

$$g(\bar{x}(t)) \leq \mathcal{D}(h;\mathcal{X})\sqrt{\frac{L}{\alpha t}}, \tag{3.22}$$

which is of the same order as the $\mathcal{O}(1/\sqrt{t})$ guarantees obtained in the references above. However, the bound (3.17) does not have a term which is antagonistic to $\eta(t)$ or $\lambda(t)$, so, if (MD) is run with constant $\lambda$ and $\eta$, we get an $\mathcal{O}(1/t)$ bound for $g(\bar{x}(t))$ (and/or $\|\bar{x}(t) - x_*\|$ in the case of strongly monotone VI problems).[6] This suggests an importang gap between continuous and discrete time; for a similar phenomenon in the context of online convex optimization, see the regret minimization analysis of [28].

We close this section with a (nonergodic) trajectory convergence result for strictly monotone problems:

**Proposition 3.5.** *Let $x_*$ denote the (necessarily unique) solution of a $\gamma$-strongly monotone problem* $\mathrm{VI}(\mathcal{X}, v)$. *If Hypotheses* (H1)–(H2) *hold and the parameters $\lambda$ and $\eta$ of* (MD) *satisfy*

$$\inf_t \lambda(t) > 0 \quad \textit{and} \quad \inf_t \eta(t) > 0, \tag{3.23}$$

*then $\lim_{t\to\infty} x(t) = x_*$ for any initialization $y(0) \in \mathcal{Y}$ of* (MD).

*Proof.* Let $\hat{x}$ be an $\omega$-limit point of $x(t)$ and assume for the purposes of obtaining a contradiction that $\hat{x} \neq x_*$. Then, by assumption, there exists an open neighborhood $O$ of $\hat{x}$ and a positive constant $a > 0$ such that

$$\langle v(x), x - x_* \rangle \geq a \quad \text{for all } x \in O. \tag{3.24}$$

Furthermore, since $\hat{x}$ is an $\omega$-limit of $x(t)$, there exists an increasing sequence $(t_k)_{k\in\mathbb{N}}$ such that $t_k \uparrow \infty$ and $x(t_k) \to \hat{x}$ as $k \to \infty$. Thus, relabeling indices if necessary, we may assume without loss of generality that $x(t_k) \in O$ for all $k \in \mathbb{N}$.

Now, for all $\varepsilon > 0$, we have

$$\begin{aligned}
\|x(t_k + \varepsilon) - x(t_k)\| &= \|Q(Y(t_k + \varepsilon)) - Q(Y(t_k))\| \\
&\leq \frac{1}{\alpha} \|Y(t_k + \varepsilon) - Y(t_k)\|_* \\
&\leq \frac{1}{\alpha} \int_{t_k}^{t_k+\varepsilon} \lambda(s) \|v(x(s))\|_* \, ds \\
&\leq \frac{1}{\alpha} \max_{x\in\mathcal{X}} \|v(x)\|_* \int_{t_k}^{t_k+\varepsilon} \lambda(s) \, ds \leq \frac{\varepsilon\bar{\lambda}}{\alpha} \max_{x\in\mathcal{X}} \|v(x)\|_*,
\end{aligned} \tag{3.25}$$

where $\bar{\lambda} = \lambda(0)$ denotes the maximum value of $\lambda(t)$. As this bound does not depend on $k$, we can choose $\varepsilon > 0$ small enough so that $x(t_k + s) \in O$ for all $s \in [0, \varepsilon]$ and all $k \in \mathbb{N}$. Thus, letting $H(t) = \eta(t)^{-1} F(x_*, \eta(t) y(t))$ and using (3.15), we obtain

$$\begin{aligned}
H(t_n) - H(t_0) &= -\sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \lambda(s) \langle v(x(s)), x(s) - x_* \rangle \, ds + \mathcal{D}(h; \mathcal{X}) \left( \frac{1}{\eta(t_n)} - \frac{1}{\eta(t_0)} \right) \\
&\leq -a\underline{\lambda} \sum_{k=1}^{n} (t_k - t_{k-1}) + \mathcal{D}(h; \mathcal{X}) \left( \frac{1}{\eta(t_n)} - \frac{1}{\eta(t_0)} \right) \\
&= -a\varepsilon\underline{\lambda} n + \mathcal{D}(h; \mathcal{X}) \left( \frac{1}{\eta(t_n)} - \frac{1}{\eta(t_0)} \right),
\end{aligned} \tag{3.26}$$

[6] In fact, even faster convergence can be guaranteed if (MD) is run with *increasing* $\lambda(t)$. In that case however, well-posedness is not immediately guaranteed, so we do not consider increasing $\lambda$ here.

where we set $\underline{\lambda} = \inf_t \lambda(t) > 0$. Given that $\inf_t \eta(t) > 0$, the above implies that $\lim_{n\to\infty} H(t_n) = -\infty$, contradicting the fact that $F(x_*, y) \geq 0$ for all $y \in \mathcal{Y}$. This implies that $\hat{x} = x_*$; by compactness, $x(t)$ admits at least one $\omega$-limit point, so our claim follows. ■

## 4. Analysis of the stochastic dynamics

In this section, we turn to the stochastic system (SMD). As in the noise-free analysis of the previous section, we begin with a well-posedness result, stated for simplicity for deterministic initial conditions of the form $Y(0) = y_0$ for a fixed $y_0 \in \mathcal{Y}$:

**Proposition 4.1.** *Fix an initial condition $y_0 \in \mathcal{Y}$. Then, under Hypotheses (H1)–(H3) and up to a $\mathbb{P}$-null set, the stochastic dynamics (SMD) admit a unique strong solution $(Y(t))_{t\geq 0}$ such that $Y(0) = y_0$.*

*Proof.* Let $B(t, y) = -\lambda(t)\sigma(Q(\eta(t)y), t)$ so (SMD) can be written as

$$dY(t) = A(t, Y(t))\, dt + B(t, Y(t))\, dW(t), \tag{4.1}$$

with $A(t, y)$ defined as in the proof of Proposition 3.1. By (H2) and (H3), $B(t, y)$ inherits the boundedness and regularity properties of $\sigma$; in particular, by Eq. (H1), it follows that $B(t, y)$ is uniformly Lipschitz in $y$. Under the same assumptions, $A(t, y)$ is also uniformly Lipschitz in $y$ (cf. the proof of Proposition 3.1). Our claim then follows by standard results in the well-posedness of stochastic differential equations [29, Theorem 3.4]. ■

We denote by $Y(t, y)$ the unique strong solution of the Itô stochastic differential equation (4.1) with initial condition $y \in \mathcal{Y}$. The corresponding primal trajectories are generated by applying the mirror map $Q$ to the dual trajectories, so that $X(t, y) = Q(\eta(t)Y(t, y))$ for all $(t, y) \in \mathbb{R}_+ \times \mathcal{Y}$. If there is no danger of confusion, we will consistently suppress the dependence on the initial condition in both random processes.

We now give a brief overview of the results we obtain in this section. First, in Section 4.1, we use the *asymptotic pseudotrajectory* (APT) theory of Benaïm and Hirsch [30] to establish almost sure trajectory convergence of (SMD) to the solution of $\mathrm{VI}(\mathcal{X}, v)$ provided that $v$ is strictly monotone and the oracle noise in (SMD) is vanishing at a rather slow, logarithmic rate. This strong convergence result relies heavily on the shadowing property of the dual trajectory and its deterministic counterpart $\phi(t, y)$. (see Section 4.1). On the other hand, if the driving noise process is persistent, we cannot expect the primal trajectory $X(t)$ to converge – some averaging has to be done in this case. Thus, following a long tradition on ergodic convergence for mirror descent, we investigate in Section 4.2 the asymptotics of a weighted time-average of $X(t)$. Finally, we complement our ergodic convergence results with a large deviation principle showing that the ergodic average of $X(t)$ is exponentially concentrated around its mean (Section 4.3).

### 4.1. The small noise limit.

We begin with the case where the oracle noise in (SMD) satisfies the asymptotic decay condition $\|\sigma(x, t)\| \leq \beta(t)$ for some nonincreasing function $\beta\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\int_0^\infty \exp\left(-\frac{c}{\beta^2(t)}\right) dt < \infty \quad \text{for all } c > 0. \tag{H4}$$

For instance, this condition is trivially satisfied if $\sigma(x,t)$ vanishes at a logarithmic rate, i.e. $\beta(t) = o(1/\sqrt{\log(t)})$. Under this decay rate condition (H4), and working for simplicity with constant $\eta(t) = \lambda(t) = 1$, the stochastic approximation theory of Benaïm and Hirsch [30, Proposition 4.1] implies that any strong solution $Y(t)$ of (SMD) is an *asymptotic pseudotrajectory* (APT) of the deterministic dynamics (MD) in the following sense:

**Definition 4.2.** Let $\phi\colon \mathbb{R}_+ \times \mathcal{Y} \to \mathcal{Y}$, $(t,y) \mapsto \phi(t,y)$, denote the semiflow induced by (MD) on $\mathcal{Y}$. A continuous curve $Y\colon \mathbb{R}_+ \to \mathcal{Y}$ is said to be an *asymptotic pseudotrajectory* (APT) of (MD) if

$$\lim_{t\to\infty} \sup_{0\le s\le T} \|Y(t+s) - \phi(s, Y(t))\|_* = 0 \quad \text{for all } T > 0. \tag{APT}$$

In words, Definition 4.2 states that an APT of (MD) tracks the solutions of (MD) to arbitrary accuracy over arbitrarily long time windows. Thanks to this property, we are able to establish the following global convergence theorem for (SMD) with vanishing oracle noise:

**Theorem 4.3.** *Assume that $v$ is strictly monotone and let $x_*$ denote the (necessarily unique) solution of* $\mathrm{VI}(\mathcal{X}, v)$. *If Hypotheses* (H1)–(H4) *hold and* (SMD) *is run with constant $\lambda(t) = \eta(t) = 1$, we have*

$$\mathbb{P}\Big(\lim_{t\to\infty} \|X(t,y) - x_*\| = 0\Big) = 1 \quad \forall y \in \mathcal{Y}. \tag{4.2}$$

The proof of Theorem 4.3 requires some auxiliary results which we provide below. We begin with a strong recurrence result for neighborhoods of the (unique) solution $x_*$ of $\mathrm{VI}(\mathcal{X}, v)$ under (MD):

**Lemma 4.4.** *With assumptions as in Theorem 4.3, let $\mathcal{O}$ be an open neighborhood of $x_*$ in $\mathcal{X}$ and define the stopping time*

$$t_{\mathcal{O}}(y) = \inf\{t \ge 0 : \xi(t,y) \in O\}, \tag{4.3}$$

*Then $t_{\mathcal{O}}(y) < \infty$ for all $y \in \mathcal{Y}$.*

*Proof.* Fix the initialization $y \in \mathcal{Y}$ of (MD), let $y(t) = \phi(t,y)$ and $x(t) = \xi(t,y) = Q(\phi(t,y))$ denote the induced solution orbit of (MD), and let $H(t) = F(x_*, y(t))$. Then, by Proposition 3.2 and the chain rule applied to (MD), we get

$$H(t) = H(0) - \int_0^t \langle v(x(s)), x(s) - x_* \rangle \, ds. \tag{4.4}$$

Since $v$ is strictly monotone and $x_*$ solves $\mathrm{VI}(\mathcal{X}, v)$, there exists some $a \equiv a_{\mathcal{O}} > 0$ such that

$$\langle v(x), x - x_* \rangle \ge a \quad \text{for all } x \in \mathcal{X} \setminus \mathcal{O}. \tag{4.5}$$

Hence, if $t_{\mathcal{O}}(y) = \infty$, we would have

$$H(t) \le H(0) - at \quad \text{for all } t \ge 0, \tag{4.6}$$

implying in turn that $\lim_{t\to\infty} H(t) = -\infty$. This contradicts the fact that $H(t) \ge 0$, so we conclude that $t_{\mathcal{O}}(y) < \infty$. ∎

Next, we extend this result to the stochastic regime:

**Lemma 4.5.** *With assumptions as in Theorem 4.3, let $\mathcal{O}$ be an open neighborhood of $x_*$ in $\mathcal{X}$ and define the stopping time*

$$\tau_{\mathcal{O}}(y) := \inf\{t \geq 0 | X(t,y) \in \mathcal{O}\}, \tag{4.7}$$

*Then, $\tau_{\mathcal{O}}(y)$ is finite (a.s.) for all $y \in \mathcal{Y}$.*

*Proof.* Suppose there exists some initial condition $y_0 \in \mathcal{Y}$ such that $\mathbb{P}(\tau_{\mathcal{O}}(y_0) = \infty) > 0$. Then there exists a measurable set $\Omega_0$ with $\mathbb{P}(\Omega_0) > 0$ and such that $\tau_{\mathcal{O}}(\omega, y_0) = \infty$ for all $\omega \in \Omega_0$. Now, define $H(t) = F(x_*, Y(t, y_0))$ and set $X(t) = X(t, y_0)$. By the weak Itô lemma (B.1) proven in Appendix B, we get

$$H(t) - H(0) \leq -\int_0^t \langle v(X(s)), X(s) - x_* \rangle \, ds + \frac{1}{2\alpha} \int_0^t \|\sigma(X(s), s)\|^2 \, ds + I_{x_*}(t) \tag{4.8}$$

where $I_{x_*}(t) = \int_0^t \langle X(s) - x_*, \sigma(X(s) \, dW(s) \rangle$ is a continuous local martingale. Since $v$ is strictly monotone, the same reasoning as in the proof of Lemma 4.4 yields

$$H(t) \leq H(0) - at + I_{x_*}(t) + \frac{\sigma_*^2}{2\alpha} \tag{4.9}$$

for some $a \equiv a_{\mathcal{O}} > 0$ and for all $t \in [0, \tau_{\mathcal{O}}(y))$. Furthermore, by an argument based on the law of the iterated logarithm and the Dambis–Dubins–Schwarz time-change theorem for martingales as in the proof of Theorem 4.6, we get

$$I_{x_*}(t)/t \to 0 \text{ almost surely as } t \to \infty. \tag{4.10}$$

Combining this with the estimate for $H(t)$ above, we get $\lim_{t\to\infty} H(t) = -\infty$ for $\mathbb{P}$-almost all $\omega \in \Omega_0$, i.e. for at least *some* $\omega \in \Omega$. This contradicts the fact that $H(t) \geq 0$ by construction, and our claim follows. ■

The above result shows that the primal process $X(t)$ hits any neighborhood of $x_*$ in finite time (a.s.). Thanks to this important recurrence property, we are finally in a position to prove Theorem 4.3:

*Proof of Theorem 4.3.* Fix some $\varepsilon > 0$ and let $N_\varepsilon = \{x = Q(y) : F(x_*, y) < \varepsilon\}$. Let $y \in \mathcal{Y}$ be arbitrary. We first claim that there exists a deterministic time $T \equiv T(\varepsilon)$ such that $F(x_*, \phi(T, y)) \leq \max\{\varepsilon, F(x_*, y) + \varepsilon\}$.

Indeed, consider the hitting time

$$t_\varepsilon(y) = \inf\{t \geq 0 : x(t) \in N_\varepsilon\}, \tag{4.11}$$

where $x(t) \equiv \xi(t, y) = Q(\phi(t, y))$. By Lemma 4.4, we have $t_\varepsilon(y) < \infty$. Moreover, observe that

$$\frac{d}{dt} F(x_*, \phi(t, y)) = -\langle v(x(t)), x(t) - x_* \rangle \leq 0 \quad \text{for all } y \in \mathcal{Y}. \tag{4.12}$$

Now, the strict monotonicity of $v$ and the fact that $x_*$ solves (VI) implies that there exists a positive constant $\kappa \equiv \kappa_\varepsilon > 0$ such that $\langle v(x), x - x_* \rangle \geq \kappa$ for all $x \in \mathcal{X} \setminus N_\varepsilon$. Hence, combining this with (4.12), we readily see that

$$F(x_*, \phi(t, y)) - F(x_*, y) \leq -\kappa t \quad \text{for all } t \in [0, t_\varepsilon(y)). \tag{4.13}$$

Now, set $T = \varepsilon/\kappa$. If $T < t_\varepsilon(y)$, we immediately conclude that

$$F(x_*, \phi(T, y)) - F(x_*, y) \leq -\varepsilon. \tag{4.14}$$

Otherwise, if $T \geq t_\varepsilon(y)$, we again use the descent property (4.12) to get

$$F(x_*, \phi(T, y)) \leq F(x_*, \phi(t_\varepsilon(y), y)) \leq \varepsilon. \tag{4.15}$$

In both cases we have $F(x_*, \phi(T, y)) \leq \max\{\varepsilon, F(x_*, y) - \varepsilon\}$, as claimed.

To proceed, pick $\delta \equiv \delta_\varepsilon > 0$ such that

$$\delta_\varepsilon \operatorname{diam}(\mathcal{X}) + \frac{\delta_\varepsilon^2}{2\alpha} < \varepsilon, \tag{4.16}$$

where $\operatorname{diam}(\mathcal{X}) = \max\{\|x' - x\|_2 : x, x' \in \mathcal{X}\}$ denotes the Euclidean diameter of $\mathcal{X}$. By Proposition 4.1 of [30], the strong solution $Y$ of (4.1) (viewed as a stochastic flow) is an APT of the deterministic semiflow $\phi$ with probability 1. Hence, we can choose an (a.s.) finite random time $\theta_\varepsilon$ such that $\sup_{s\in[0,T]} \|Y(t+s) - \phi(s, Y(t))\|_* \leq \delta_\varepsilon$ for all $t \geq \theta_\varepsilon$. Combining this with item (c) of Proposition 3.2, we then get

$$\begin{aligned} F(x_*, Y(t+s, y)) &\leq F(x_*, \phi(s, Y(t, y))) \\ &\quad + \langle Y(t+s, y) - \phi(s, Y(t, y)), Q(\phi(s, Y(t, y))) - x_* \rangle \\ &\quad + \frac{1}{2\alpha} \|Y(t+s, y) - \phi(s, Y(t, y))\|_*^2 \\ &\leq F(x_*, \phi(s, Y(t, y))) + \delta_\varepsilon \operatorname{diam}(\mathcal{X}) + \frac{\delta_\varepsilon^2}{2\alpha} \\ &\leq F(x_*, \phi(s, Y(t, y))) + \varepsilon, \end{aligned} \tag{4.17}$$

where the last inequality follows from the estimate (4.16).

Now, choose a random time $T_0 \geq \max\{\theta_\varepsilon(y), t_\varepsilon(y)\}$ and $T = \varepsilon/\kappa$ as above. Then, by definition, we have $F(x_*, Y(T_0, y)) \leq 2\varepsilon$ with probability 1. Hence, for all $s \in [0, T]$, we get

$$F(x_*, Y(T_0 + s, y)) \leq F(x_*, \phi(s, Y(T_0, y))) + \varepsilon \leq F(x_*, Y(T_0, y)) + \varepsilon \leq 3\varepsilon. \tag{4.18}$$

Since $F(x_*, \phi(T, Y(T_0, y))) \leq \max\{\varepsilon, F(x_*, Y(T_0, y)) - \varepsilon\} \leq \varepsilon$, we also get

$$\begin{aligned} F(x_*, Y(T_0 + T + s, y)) &\leq F(x_*, \phi(s, Y(T_0 + T, y))) + \varepsilon \\ &\leq F(x_*, Y(T_0 + T, y)) + \varepsilon \\ &\leq 3\varepsilon, \end{aligned} \tag{4.19}$$

and hence

$$F(x_*, Y(T_0 + s, y)) \leq 3\varepsilon \quad \text{for all } s \in [T, 2T]. \tag{4.20}$$

Using this as the basis for an induction argument, we readily get

$$F(x_*, Y(T_0 + s, y)) \leq 3\varepsilon \quad \text{for all } s \in [nT, (n+1)T], \tag{4.21}$$

with probability 1. Since $\varepsilon$ was arbitrary, we obtain $F(x_*, Y(t, y)) \to 0$, implying in turn that $X(t) \to x_*$ (a.s.) by Proposition 3.2. ■

4.2. **Ergodic Convergence.** We now proceed with an ergodic convergence result in the spirit of Proposition 3.3. To state it, set

$$S(t) = \int_0^t \lambda(s)\, ds \quad \text{and} \quad L(t) = \sqrt{\int_0^t \lambda^2(s)\, ds}, \tag{4.22}$$

and let

$$\bar{X}(t) = \frac{1}{S(t)} \int_0^t \lambda(s) X(s)\, ds, \tag{4.23}$$

denote the "ergodic average" of $X(t)$. Our main result may then be stated as follows:

**Theorem 4.6.** *Under Hypotheses (H1)–(H3), we have:*

$$g(\bar{X}(t)) = \mathcal{O}\left(\frac{1}{\eta(t)S(t)}\right) + \mathcal{O}\left(\frac{\int_0^t \lambda^2(s)\eta(s)\,ds}{S(t)}\right) + \mathcal{O}\left(\frac{L(t)\sqrt{\log\log L(t)}}{S(t)}\right), \quad (4.24)$$

*with probability* 1. *In particular,* $\bar{X}(t)$ *converges (a.s.) to the solution set of* $\mathrm{VI}(\mathcal{X}, v)$ *provided that a)* $\lim_{t\to\infty} \eta(t)S(t) = \infty$*; and b)* $\lim_{t\to\infty} \eta(t)\lambda(t) = 0$.

The proof of Theorem 4.6 relies crucially on the following lemma, which provides an explicit estimate for the decay rate of $g(\bar{X}(t))$, both for generic VI problems and convex-concave saddle-point problems:

**Lemma 4.7.** *If (SMD) is initialized at* $y_0 = 0$ *and Hypotheses (H1)–(H3) hold, then:*

$$g(\bar{X}(t)) \leq \frac{\mathcal{D}(h;\mathcal{X})}{\eta(t)S(t)} + \frac{\sigma_*^2}{2\alpha}\frac{\int_0^t \lambda^2(s)\eta(s)\,ds}{S(t)} + \frac{I(t)}{S(t)} \quad (4.25)$$

*where* $I(t) = \sup_{p\in\mathcal{X}} I_p(t)$ *and*

$$I_p(t) = \int_0^t \lambda(s)\langle p - X(s), \sigma(X(s),s)\cdot dW(s)\rangle. \quad (4.26)$$

*In particular, if (VI) is associated with a convex-concave saddle-point problem as in Example 2.2, we have*

$$G(\bar{X}^1(t), \bar{X}^2(t)) \leq \frac{\mathcal{D}_{\mathrm{sp}}}{\eta(t)S(t)} + \frac{\sigma_*^2}{2\alpha_{\mathrm{sp}}}\frac{\int_0^t \lambda^2(s)\eta(s)\,ds}{S(t)} + \frac{J(t)}{S(t)}, \quad (4.27)$$

*where we have set* $\mathcal{D}_{\mathrm{sp}} = \mathcal{D}(h_1;\mathcal{X}^1) + \mathcal{D}(h_2;\mathcal{X}^2)$, $1/\alpha_{\mathrm{sp}} = 1/\alpha_1 + 1/\alpha_2$, *and* $J(t) = \sup_{p^1\in\mathcal{X}^1, p^2\in\mathcal{X}^2}\{I_{p^1}(t) + I_{p^2}(t)\}$.

*Remark* 4.1. The initialization assumption in Lemma 4.7 is not crucial: we only make it to simplify the explicit expression (4.25). If (SMD) is initialized at a different point, the proof of Lemma 4.7 shows that the bound (4.25) is correct only up to $\mathcal{O}(1/S(t))$. Since all terms in (4.25) are no faster than $\mathcal{O}(1/S(t))$, initialization plays no role in the proof of Theorem 4.6 below.

*Proof of Lemma 4.7.* Fix some $p \in \mathcal{X}$ and let $H_p(t) = \eta(t)^{-1}F(p, \eta(t)Y(t))$ as in the proof of Proposition 3.3. Then, by the weak Itô formula (B.1) in Appendix B, we have

$$\begin{aligned} H_p(t) \leq H_p(0) &- \int_0^t \frac{\dot{\eta}(s)}{\eta(s)^2}H_p(s)\,ds + \frac{1}{\eta(t)}\int_0^t \langle X(s) - p, \dot{\eta}(s)Y(s)\rangle\,ds \\ &+ \int_0^t \langle X(s) - p,\, dY(s)\rangle + \frac{1}{2\alpha}\int_0^t \lambda^2(s)\eta(s)\|\sigma(X(s),s)\|^2\,ds. \end{aligned} \quad (4.28)$$

To proceed, let

$$R_p(t) = \int_0^t \lambda(s)\langle v(X(s)), X(s) - p\rangle\,ds, \quad (4.29)$$

so $\int_0^t \langle X(s) - p, dY(s)\rangle = -\int_0^t \lambda(s)\langle X(s) - p, v(X(s))\,ds + dM(s)\rangle = -R_p(t) + I_p(t)$, with $I_p(t)$ given by (4.26). Then, rearranging and bounding the second term of (4.28) as in the proof of Proposition 3.3, we obtain

$$R_p(t) \leq H_p(0) - H_p(t) + \mathcal{D}(h,\mathcal{X})\left(\frac{1}{\eta(t)} - \frac{1}{\eta(0)}\right)$$

$$
\begin{aligned}
&+ I_p(t) + \frac{1}{2\alpha}\int_0^t \lambda^2(s)\eta(s)\|\sigma(X(s),s)\|^2\,ds \\
&\leq H_p(0) + \mathcal{D}(h;\mathcal{X})\left(\frac{1}{\eta(t)} - \frac{1}{\eta(0)}\right) + I_p(t) + \frac{\sigma_*^2}{2\alpha}\int_0^t \lambda^2(s)\eta(s)\,ds. \qquad (4.30)
\end{aligned}
$$

With (SMD) initialized at $y_0 = 0$, Eq. (3.16) gives $H_p(0) \leq \mathcal{D}(h;\mathcal{X})/\eta(0)$. Thus, by Jensen's inequality and the monotonicity of $v$, we get

$$
\begin{aligned}
\langle v(p), \bar{X}(t) - p\rangle &= \frac{1}{S(t)}\int_0^t \lambda(s)\langle v(p), X(s) - p\rangle\,ds \\
&\leq \frac{1}{S(t)}\int_0^t \lambda(s)\langle v(X(s)), X(s) - p\rangle\,ds = \frac{R_p(t)}{S(t)} \\
&\leq \frac{\mathcal{D}(h;\mathcal{X})}{\eta(t)S(t)} + \frac{\sigma_*^2}{2\alpha}\frac{\int_0^t \lambda^2(s)\eta(s)\,ds}{S(t)} + \frac{I_p(t)}{S(t)}. \qquad (4.31)
\end{aligned}
$$

The bound (4.25) then follows by noting that $g(\bar{X}(t)) = \max_{p\in\mathcal{X}}\langle v(p), \bar{X}(t) - p\rangle$.

Now, assume that (VI) is associated to a convex-concave saddle point problem as in Example 2.2. As in the proof of Proposition 3.3, we first replicate the analysis above for each component of the problem, and we then sum the two components to get an overall bound for the Nikaido–Isoda gap function $G$. Specifically, applying (4.31) to (2.2), we readily get

$$
\int_0^t \lambda(s)\langle v^i(X(s)), X^i(s) - p^i\rangle \leq \frac{\mathcal{D}(h_i;\mathcal{X}^i)}{\eta(t)S(t)} + \frac{\sigma_*^2}{2\alpha^i}\frac{\int_0^t \lambda^2(s)\eta(s)\,ds}{S(t)} + \frac{I_{p^i}(t)}{S(t)}, \qquad (4.32)
$$

where $i \in \{1,2\}$. Moreover, combining Jensen's inequality with the fact that $U$ is convex-concave yields

$$
\begin{aligned}
U(\bar{X}^1(t), p^2) - U(p^1, \bar{X}^2(t)) &\leq \frac{1}{S(t)}\int_0^t \lambda(s)[U(X^1(s), p^2) - U(p^1, X^2(s))]\,ds \\
&\leq \frac{1}{S(t)}\int_0^t \lambda(s)\langle \nabla_{x^1}U(X(s)), X^1(s) - p^1\rangle\,ds \\
&\quad - \frac{1}{S(t)}\int_0^t \lambda(s)\langle \nabla_{x^2}U(X(s)), X^2(s) - p^2\rangle\,ds \\
&\leq \frac{\mathcal{D}_{\mathrm{sp}}}{\eta(t)S(t)} + \frac{\sigma_*^2}{2\alpha_{\mathrm{sp}}}\frac{\int_0^t \lambda^2(s)\eta(s)\,ds}{S(t)} + \frac{I_{p^1}(t) + I_{p^2}(t)}{S(t)}, \qquad (4.33)
\end{aligned}
$$

with the last ineqality following from (4.32). Our claim then follows by maximizing over $(p^1, p^2)$ and recalling the definition (3.3) of the Nikaido–Isoda gap function. ■

Clearly, the crucial unknown in the bound (4.25) is the stochastic term $I(t)$: to obtain convergence of $\bar{X}(t)$ to the solution set of $\mathrm{VI}(\mathcal{X}, v)$, the term $I(t)$ must grow slower than $S(t)$. As we show below, this is indeed the case:

*Proof of Theorem 4.6.* By Lemma 4.7 and Remark 4.1, it suffices to show that the term $I(t)$ grows as $\mathcal{O}(L(t)\sqrt{\log\log L(t)})$ with probability 1. To do so, let $\phi_p = [I_p]$

denote the quadratic variation of $I_p$;[7] then, the rules of stochastic calculus yield

$$\begin{aligned} d\phi_p(t) &= dI_p(t) \cdot dI_p(t) \\ &= \lambda^2(t) \sum_{i,j=1}^{n} \sum_{k=1}^{d} (X_i(t) - p_i)(X_j(t) - p_j)\sigma_{ik}(X(t),t)\sigma_{jk}(X(t),t)\, dt \\ &\leq \|X(t) - p\|_2^2 \sigma_*^2 \lambda^2(t) \leq \operatorname{diam}(\mathcal{X})^2 \sigma_*^2 \lambda^2(t), \end{aligned} \tag{4.34}$$

where $\operatorname{diam}(\mathcal{X}) = \max\{\|x' - x\|_2 : x, x' \in \mathcal{X}\}$ denotes the Euclidean diameter of $\mathcal{X}$. Hence, for all $t \geq 0$, we get the covariation bound

$$\phi_p(t) \leq \operatorname{diam}(\mathcal{X})^2 \sigma_*^2 \int_0^t \lambda^2(s)\, ds = \mathcal{O}(L^2(t)). \tag{4.35}$$

Now, let $\phi_p(\infty) = \lim_{t\to\infty} \phi_p(t) \in [0, \infty]$ and set

$$\tau_p(s) = \begin{cases} \inf\{t \geq 0 : \phi(t) > s\} & \text{if } s \leq \phi_p(\infty), \\ \infty & \text{otherwise.} \end{cases} \tag{4.36}$$

The process $\tau_p(s)$ is finite, non-negative, non-decreasing and right-continuous on $[0, \phi_p(\infty))$; moreover, it is easy to check that $\phi_p(\tau_p(s)) = s \wedge \phi_p(\infty)$ and $\tau_p(\phi_p(t)) = t$ [31, Problem 3.4.5]. Therefore, by the Dambis–Dubins–Schwarz time-change theorem for martingales [31, Thm. 3.4.6 and Pb. 3.4.7], there exists a standard, one-dimensional Wiener process $(B_p(t))_{t\geq 0}$ adapted to a modified filtration $\tilde{\mathcal{F}}_s = \mathcal{F}_{\tau_p(s)}$ (possibly defined on an extended probability space), and such that $B_p(\phi_p(t)) = I_p(t)$ for all $t \geq 0$ (except possibly on a $\mathbb{P}$-null set).

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

$$\frac{I_p(t)}{S(t)} = \frac{B_p(\phi_p(t))}{S(t)} = \frac{B_p(\phi_p(t))}{\sqrt{\phi_p(t) \log\log \phi_p(t)}} \times \frac{\sqrt{\phi_p(t) \log\log \phi_p(t)}}{S(t)}. \tag{4.37}$$

By the law of the iterated logarithm [31], the first factor above is bounded almost surely; as for the second, (4.35) gives $\sqrt{\phi_p(t) \log\log \phi_p(t)} = \mathcal{O}(L(t)\sqrt{\log\log L(t)})$. Thus, combining all of the above, we get

$$\frac{I(t)}{S(t)} = \frac{\max_{p\in\mathcal{X}} I_p(t)}{S(t)} = \mathcal{O}\left(\frac{L(t)\sqrt{\log\log L(t)}}{S(t)}\right), \tag{4.38}$$

so our claim follows from (4.25).

To complete our proof, note first that the condition $\lim_{t\to\infty} \eta(t)S(t) = \infty$ implies that $\lim_{t\to\infty} S(t) = \infty$ (given that $\eta(t)$ is nonincreasing). Thus, by de l'Hôpital's rule and the assumption $\lim_{t\to\infty} \lambda(t)\eta(t) = 0$, we also get $S(t)^{-1} \int_0^t \lambda^2(s)\eta(s)ds = 0$. Finally, for the last term of (4.24), consider the following two cases:

(1) If $\lim_{t\to\infty} L(t) < \infty$, we trivially have $\lim_{t\to\infty} L(t)\sqrt{\log\log L(t)}/S(t) = 0$ as well.
(2) Otherwise, if $\lim_{t\to\infty} L(t) = \infty$, de l'Hôpital's rule readily yields

$$\lim_{t\to\infty} \frac{L^2(t)}{S^2(t)} = \lim_{t\to\infty} \frac{\lambda^2(t)}{2\lambda(t)S(t)} = \frac{1}{2} \lim_{t\to\infty} \frac{\lambda(t)}{S(t)} = 0, \tag{4.39}$$

[7] Recall here that the quadratic variation of a stochastic process $M(t)$ is defined as $[M(t)] = \lim_{|\Pi|\to 0} \sum_{1\leq j\leq k} (M(t_j) - M(t_{j-1}))^2$, where the limit is taken over all partitions $\Pi = \{t_0 = 0 < t_1 < \cdots < t_k = t\}$ of $[0, t]$ with mesh $|\Pi| \equiv \max_j |t_j - t_{j-1}| \to 0$ [31].

by the boundedness of $\lambda(t)$. Another application of de l'Hôpital's rule gives

$$\lim_{t\to\infty} \frac{L^3(t)}{S^2(t)} = \lim_{t\to\infty} \frac{(L^2(t))^{3/2}}{S^2(t)} = \frac{3}{4} \lim_{t\to\infty} \frac{\lambda^2(t)L(t)}{\lambda(t)S(t)} = \frac{3}{4} \lim_{t\to\infty} \frac{\lambda(t)L(t)}{S(t)} = 0, \tag{4.40}$$

so

$$\limsup_{t\to\infty} \frac{L(t)\sqrt{\log\log L(t)}}{S(t)} \leq \limsup_{t\to\infty} \sqrt{\frac{L^3(t)}{S^2(t)}} = 0. \tag{4.41}$$

The above shows that, under the stated assumptions, the RHS of (4.24) converges to 0 (a.s.), implying in turn that $\bar{X}(t)$ converges to the solution set of $\mathrm{VI}(\mathcal{X}, v)$ with probability 1. ■

4.3. **Large deviations.** In this section we study the concentration properties of (SMD) in terms on the dual gap function. First, recall that for every $p \in \mathcal{X}$ we have the upper bound

$$R_p(t) \leq \frac{\mathcal{D}(h;\mathcal{X})}{\eta(t)} + \frac{\sigma_*^2}{2\alpha} \int_0^t \lambda^2(s)\eta(s)\, ds + I_p(t). \tag{4.42}$$

with $R_p(t)$ and $I_p(t)$ defined as in (4.29) and (4.26) respectively. Since $I_p(t)$ is a continuous martingale starting at 0, we have $\mathbb{E}[I_p(t)] = 0$, implying in turn that

$$\mathbb{E}[\langle v(p), \bar{X}(t) - p\rangle] \leq \frac{\mathcal{D}(h;\mathcal{X})}{S(t)\eta(t)} + \frac{\sigma_*^2}{2\alpha S(t)} \int_0^t \lambda^2(s)\eta(s)\, ds = \frac{K(t)}{2S(t)}, \tag{4.43}$$

where

$$K(t) = \frac{2\mathcal{D}(h;\mathcal{X})}{\eta(t)} + \frac{\sigma_*^2}{\alpha} \int_0^t \lambda^2(s)\eta(s)\, ds. \tag{4.44}$$

Hence, taking the supremum over all $p \in \mathcal{X}$ and using Jensen's inequality, we get the mean gap bound

$$\mathbb{E}[g(\bar{X}(t)] \leq \frac{K(t)}{2S(t)}. \tag{4.45}$$

Markov's inequality therefore implies that

$$\mathbb{P}(g(\bar{X}(t)) \geq \delta) \leq \frac{1}{\delta}\frac{K(t)}{2S(t)} \quad \text{for all } \delta > 0. \tag{4.46}$$

The bound (4.46) provides a first estimate of the probability of observing a large gap from the solution of (VI), but because it relies only on Markov's inequality, it is rather crude. To refine it, we provide below a "large deviations" bound that shows that the ergodic gap process $g(\bar{X}(t))$ is exponentially concentrated around its mean value:

**Theorem 4.8.** *Suppose* (H1)–(H3) *hold. Then, for all $\delta > 0$ and all $t > 0$, we have*

$$\mathbb{P}(g(\bar{X}(t)) \geq \mathcal{Q}_0(t) + \delta\mathcal{Q}_1(t)) \leq \exp(-\delta^2/4), \tag{4.47}$$

*where*

$$\mathcal{Q}_0(t) = \frac{K(t)}{S(t)}, \tag{4.48a}$$

*and*

$$\mathcal{Q}_1(t) = \frac{\sqrt{\kappa}\sigma_* \operatorname{diam}(\mathcal{X})L(t)}{S(t)}, \tag{4.48b}$$

*with $\kappa > 0$ a positive constant depending only on the set $\mathcal{X}$ and the norm $\|\cdot\|$.*

To prove Theorem 4.8 we need some groundwork first. To that end, define the auxiliary processes

$$Z(t) = \int_0^t \lambda(s)\sigma(X(s), s)\, dW(s), \tag{4.49}$$

and

$$P(t) = Q(\eta(t)Z(t)). \tag{4.50}$$

We then have:

**Lemma 4.9.** *For all $p \in \mathcal{X}$ we have*

$$\int_0^t \lambda(s)\langle p - P(s), \sigma(X(s), s)\, dW(s)\rangle \leq \frac{\mathcal{D}(h;\mathcal{X})}{\eta(t)} + \frac{\sigma_*^2}{2\alpha}\int_0^t \lambda^2(s)\eta(s)\, ds. \tag{4.51}$$

*Proof.* The proof follows the same lines as Lemma 4.7. Specifcially, given a reference point $p \in \mathcal{X}$, define the process $\tilde{H}_p(t) = \frac{1}{\eta(t)}F(p, \eta(t)Z(t))$. Then, by the weak Itô formula (B.1) in Appendix B, we have

$$\begin{aligned}
\tilde{H}_p(t) &\leq \tilde{H}_p(0) - \int_0^t \frac{\dot{\eta}(s)}{\eta(s)^2}\tilde{H}_p(s)\, ds + \frac{1}{\eta(t)}\int_0^t \langle \xi(s) - p, \dot{\eta}(s)Z(s)\rangle\, ds \\
&\quad + \int_0^t \langle P(s) - p,\, dZ(s)\rangle + \frac{1}{2\alpha}\int_0^t \lambda^2(s)\eta(s)\|\sigma(X(s), s)\|^2\, ds \\
&\leq -\int_0^t \frac{\dot{\eta}(s)}{\eta(s)}[h(p) - h(P(s))]\, ds \\
&\quad + \int_0^t \lambda(s)\langle P(s) - p, \sigma(X(s), s)\, dW(s)\rangle + \frac{\sigma_*^2}{2\alpha}\int_0^t \lambda^2(s)\eta(s)\, ds.
\end{aligned} \tag{4.52}$$

We thus get,

$$\begin{aligned}
\int_0^t \lambda(s)\langle p - P(s), \sigma(X(s), s)\, dW(s)\rangle &\leq \tilde{H}_p(0) + \mathcal{D}(h;\mathcal{X})\left(\frac{1}{\eta(t)} - \frac{1}{\eta(0)}\right) + \int_0^t \frac{\lambda^2(s)\eta(s)}{2\alpha}\sigma_*^2\, ds \\
&\leq \frac{\mathcal{D}(h;\mathcal{X})}{\eta(t)} + \int_0^t \frac{\lambda^2(s)\eta(s)}{2\alpha}\sigma_*^2\, ds,
\end{aligned} \tag{4.53}$$

as claimed. ■

We are now ready to establish our large deviations principle for (SMD):

*Proof of Theorem 4.8.* For $p \in \mathcal{X}$ and $t > 0$ fixed, we have

$$\begin{aligned}
R_p(t) &\leq \frac{\mathcal{D}(h;\mathcal{X})}{\eta(t)} + \frac{\sigma_*^2}{2\alpha}\int_0^t \lambda^2(s)\eta(s)\, ds + \int_0^t \lambda(s)\langle p - X(s), \sigma(X(s), s)\, dW(s)\rangle \\
&= \frac{\mathcal{D}(h;\mathcal{X})}{\eta(t)} + \frac{\sigma_*^2}{2\alpha}\int_0^t \lambda^2(s)\eta(s)\, ds + \int_0^t \lambda(s)\langle p - P(s), \sigma(X(s), s)\, dW(s)\rangle \\
&\quad + \int_0^t \lambda(s)\langle P(s) - X(s), \sigma(X(s), s)\, dW(s)\rangle \\
&\leq \frac{2\mathcal{D}(h;\mathcal{X})}{\eta(t)} + \frac{\sigma_*^2}{\alpha}\int_0^t \lambda^2(s)\eta(s)\, ds + \int_0^t \langle P(s) - X(s), \sigma(X(s), s)\, dW(s)\rangle,
\end{aligned} \tag{4.54}$$

where we used (4.51) to obtain the last inequality. Now, to proceed, let

$$\Delta(t) = \int_0^t \lambda(s)\langle P(s) - X(s), \sigma(X(s), s)\, dW(s)\rangle. \tag{4.55}$$

The process $\Delta(t)$ is a continuous martingale starting at 0 which is almost surely bounded in $L^2$. In particular, it allows us to give an upper bound on $R_p(t)$ which is independent of the reference point $p \in \mathcal{X}$. Indeed, recalling that $K(t) = 2\mathcal{D}(h;\mathcal{X})/\eta(t) + \alpha^{-1}\sigma_*^2 \int_0^t \lambda^2(s)\eta(s)\, ds$, we see that

$$R_p(t) \leq K(t) + \Delta(t), \tag{4.56}$$

and hence

$$g(\bar{X}(t)) \leq \frac{K(t) + \Delta(t)}{S(t)} \qquad \forall t > 0. \tag{4.57}$$

This implies that for all $\varepsilon, t > 0$,

$$\{g(\bar{X}(t) \geq \varepsilon\} \subseteq \{\Delta(t) \geq \varepsilon S(t) - K(t)\},$$

and hence

$$\mathbb{P}[g(\bar{X}(t)) \geq \varepsilon] \leq \mathbb{P}[\Delta(t) \geq \varepsilon S(t) - K(t)].$$

It remains to bound the right-hand side of the above expression. To this end, observe that for all $b, \theta > 0$, the Schwarz inequality gives us

$$\begin{aligned}\mathbb{E}[\exp(\theta\Delta(t))] &= \mathbb{E}[\exp(\theta\Delta(t) - b\langle\Delta\rangle_t)\exp(b\langle\Delta\rangle_t)] \\ &\leq \sqrt{\mathbb{E}[\exp(2\theta\Delta(t) - 2b\langle\Delta\rangle_t)]}\sqrt{\mathbb{E}[\exp(2b\langle\Delta\rangle_t)]}.\end{aligned}$$

Setting $b = \theta^2$, the expressions inside the first expected value is just the stochastic exponential of the process $(\theta\Delta(t))_{t\geq 0}$. A straightforward computations shows that

$$\begin{aligned}\langle\Delta\rangle_t &= \int_0^t \lambda^2(s)\|\sigma(X(s), s)^\top(P(s) - X(s))\|_2^2\, ds \\ &\leq \kappa \int_0^t \lambda^2(s)\|\sigma(X(s), s)\|^2\|P(s) - X(s)\|^2\, ds \\ &\leq \kappa\sigma_*^2 \operatorname{diam}(\mathcal{X})^2 L(t)^2 \\ &=: \varphi(t).\end{aligned}$$

The constant $\kappa$ is a universal factor accounting for the equivalence of the Euclidean norm and the norm $\|\cdot\|$ on $\mathcal{X}$. Given the a.s. boundedness of the quadratic variation on each compact interval, the process $\{(\Delta(t), \mathcal{F}_t); t \geq 0\}$ is a true martingale with expected value 1. Hence,

$$\mathbb{E}[\exp(\theta\Delta(t))] \leq \sqrt{\mathbb{E}[\exp(2\theta^2\langle\Delta\rangle_t)]} \leq \exp(\theta^2\varphi(t)). \tag{4.58}$$

Using all these facts, we see that for all $\Omega > 0$

$$\begin{aligned}\mathbb{P}[\Delta(t) \geq \Omega] &= \mathbb{P}[\exp(\theta\Delta(t)) \geq \exp(\theta\Omega)] \\ &\leq \exp(-\theta\Omega)\mathbb{E}[\exp(\theta\Delta(t))] \quad \text{(Markov Inequality)} \\ &= \exp(-\theta\Omega + \theta^2\varphi(t)) \quad (4.58).\end{aligned}$$

Minimizing this with respect to $\theta$ gives for all $t > 0$ and $\Omega > 0$

$$\mathbb{P}[\Delta(t) \geq \Omega] \leq \exp\left(-\frac{\Omega^2}{4\varphi(t)}\right).$$

If we introduce the functions (4.48a)-(4.48b), we finally arrive at the expression

$$\begin{aligned}\mathbb{P}(g(\bar{X}(t)) \geq \mathcal{Q}_0(t) + \delta\mathcal{Q}_1(t)) &\leq \mathbb{P}(\Delta(t) \geq \mathcal{Q}_0(t)S(t) + \delta\mathcal{Q}_1(t)S(t) - K(t)) \\ &\leq \mathbb{P}(\Delta(t) \geq \delta\sqrt{\varphi(t)}) \\ &\leq \exp(-\delta^2/4), \end{aligned} \tag{4.59}$$

as claimed. ■

## 5. Conclusion

This paper examined a continuous-time dynamical system for solving monotone variational inequality problems with random inputs. The key element of our analysis is the identification of a energy-type function, which allows us to prove ergodic convergence of generated trajectories in the deterministic as well as in the stochastic case. Future research should extend the present work in the following dimensions. First, it is not clear yet how the continuous-time method will help us in the derivation of a consistent numerical scheme. A naive Euler-discretization might potentially lead to a loss in speed of convergence (see [2]). Second, it is of great interest to relax the monotonicity assumption we made on the involved operator. We are currently investigating these extensions.

## Appendix A. Estimates

In this appendix we collect some simple facts on the analysis of convex differentiable functions with Lipschitz continuous gradients. Denote by $\mathbf{C}^{1,1}_L(\mathbb{R}^n)$ the totality of such functions, with $L$ being the Lipschitz constant of the gradient mapping $\nabla\psi$:

$$\|\psi(y+\delta) - \psi(y)\|_* \leq L\|\delta\|_* \qquad \forall y, \delta \in \mathbb{R}^n.$$

**Proposition A.1.** *Let $\psi \in \mathbf{C}^{1,1}_L(\mathbb{R}^n)$ be convex. Then $\psi$ is almost everywhere twice differentiable with Hessian $\nabla^2\psi$ and*

$$0 \leq \nabla^2\psi(y) \leq L\,\mathrm{Id}\,. \quad \mathsf{Leb} - a.e.. \tag{A.1}$$

*Proof.* For every $\psi \in \mathbf{C}^{1,1}_L(\mathbb{R}^n)$, the well-known descent lemma ([32], Theorem 2.1.5) implies that

$$\psi(y+\delta) \leq \psi(y) + \langle\nabla\psi(y), \delta\rangle + \frac{L}{2}\|\delta\|_*^2 \qquad \forall y, \delta \in \mathbb{R}^n. \tag{A.2}$$

By Alexandrov's theorem (see e.g. [33], Lemma 6.6), it follows that $\psi$ is $\mathsf{Leb}$-almost everywhere twice differentiable. Hence, there exists a measurable set $\Lambda$ such that $\mathsf{Leb}(\Lambda) = 0$, and for all $\bar{y} \in \mathbb{R}^n \setminus \Lambda$ there exists $(p, P) \in \mathbb{R}^n \times \mathbb{R}^{n\times n}_{sym}$ such that

$$\psi(\bar{y}+y) = \psi(\bar{y}) + \langle p, y\rangle + \frac{1}{2}\langle Py, y\rangle + \theta(\bar{y}, y), \tag{A.3}$$

where $\lim_{\|y\|_*\to 0}\frac{\theta(\bar{y},y)}{\|y\|_*^2} = 0$. We have $p = \nabla\psi(\bar{y})$ and identify $P$ with the a.e. defined Hessian $\nabla^2\psi(\bar{y})$. On the other hand, convexity implies

$$\psi(\bar{y}+y) \geq \psi(\bar{y}) + \langle\nabla\psi(\bar{y}), y\rangle \qquad \forall \bar{y}, y \in \mathbb{R}^n. \tag{A.4}$$

Choosing $y = te$, where $e \in \mathbb{R}^n$ is an arbitrary $\|\cdot\|_*$-unit vector and $t > 0$, it follows

$$-\frac{1}{t^2}\theta(\bar{y}, te) \overset{(A.4)}{\leq} \frac{1}{t^2}[\psi(\bar{y}+te) - \psi(\bar{y}) - \langle\nabla\psi(\bar{y}), te\rangle] - \frac{1}{t^2}\theta(\bar{y}, te)$$

$$\overset{(A.3)}{=} \frac{1}{2}\langle \nabla^2\psi(\bar{y})e, e\rangle$$
$$\overset{(A.2)}{\leq} \frac{L}{2} - \frac{1}{t^2}\theta(\bar{y}, te).$$

Letting $t \to 0^+$ we get

$$0 \leq \frac{1}{2}\langle \nabla^2\psi(\bar{y})e, e\rangle \leq \frac{L}{2} \qquad \forall \bar{y} \in \mathbb{R}^n \setminus \Lambda, \tag{A.5}$$

which implies $\nabla^2\psi(\bar{y}) \leq L\,\mathrm{Id}$. ■

## Appendix B. Results from stochastic analysis

The following result is the generalized Itô formula used in the main text.

**Proposition B.1.** *Let $Y$ be an Itô process in $\mathbb{R}^n$ of the form*

$$Y_t = Y_0 + \int_0^t F_s\, ds + \int_0^t G_s\, dW(s).$$

*Let $\psi \in \mathbf{C}_L^{1,1}(\mathbb{R}^n)$ be convex. Then for all $t \geq 0$ we have*

$$\psi(Y_t) \leq \psi(Y_0) + \int_0^t \langle \nabla\psi(Y_s),\, dY_s\rangle + \frac{L}{2}\int_0^t \|G_s\|^2\, ds$$

*Proof.* Since $\psi \in \mathbf{C}_L^{1,1}(\mathbb{R}^n)$ is convex, Proposition A.1 shows that $\psi$ is almost everywhere twice differentiable with Hessian $\nabla^2\psi$. Furthermore, this Hessian matrix satisfies $0 \leq \nabla^2\psi(y) \leq L\,\mathrm{Id}$, for all $y \in \mathbb{R}^n$ outside a set of Lebesgue measure 0. Introduce the mollifier

$$\rho(u) := \begin{cases} c\exp\left(\frac{-1}{1-\|u\|_*^2}\right) & \text{if } \|u\|_* < 1, \\ 0 & \text{if } \|u\|_* \geq 1. \end{cases}$$

Choose the constant $c > 0$ so that $\int_{\mathbb{R}^n} \rho(u)\, du = 1$. For every $\varepsilon > 0$ define

$$\rho_\varepsilon(u) = \varepsilon^{-n}\rho(u/\varepsilon),$$
$$\psi_\varepsilon(y) = \psi \circledast \rho_\varepsilon(y) := \int_{\mathbb{R}^n} \psi(y-u)\rho_\varepsilon(u)\, du.$$

Then $\psi_\varepsilon \in \mathbf{C}^\infty(\mathbb{R}^n)$ and the standard form of Itô's formula gives us

$$\begin{aligned}\psi_\varepsilon(Y_t) &= \psi_\varepsilon(Y_s) + \int_s^t \langle \nabla\psi_\varepsilon(Y_r),\, dY_r\rangle + \frac{1}{2}\int_s^t \mathrm{tr}\left[\nabla^2\psi_\varepsilon(Y_r)G_rG_r^\top\right]\, dr \\ &= \psi_\varepsilon(Y_s) + \int_s^t \langle \int_{\mathbb{R}^n} \nabla\psi(z)\rho_\varepsilon(Y_r - z)\, dz,\, dY_r\rangle \\ &\quad + \frac{1}{2}\int_s^t \int_{\mathbb{R}^n} \mathrm{tr}\left[\nabla^2\psi(z)G_rG_r^\top\right]\rho_\varepsilon(Y_r - z)\, dr\, dz.\end{aligned}$$

Since $\mathrm{tr}(\nabla^2\psi(z)G_rG_r^\top) \leq L\|G_r\|^2$, we get

$$\psi_\varepsilon(Y_t) \leq \psi_\varepsilon(Y_s) + \int_s^t \langle \int_{\mathbb{R}^n} \nabla\psi(z)\rho_\varepsilon(Y_r - z)\, dz,\, dY_r\rangle + \frac{L}{2}\int_s^t \|G_r\|^2\, dr.$$

Letting $\varepsilon \to 0^+$, using the uniform convergence of the involved data, proves the result. ■

*Remark* B.1. Applying this result to the dual process of (SMD) and using (A.5), gives for $F_s = -\lambda(s)v(X_s)$ and $G_s = -\lambda(s)\sigma(X_s, s)$ the following version of the generalized Itô rule:

$$\psi(Y_t) \leq \psi(Y_0) - \int_0^t \langle \nabla\psi(Y_s),\, dY_s\rangle\, ds + \frac{1}{2\alpha}\int_0^t \|\sigma(X_s, s)\|^2\, ds \tag{B.1}$$

## References


[1] Peypouquet, J., Sorin, S.: Evolution equations for maximal monotone operators: Asymptotic analysis in continuous and discrete time. Journal of Convex Analysis pp. 1113–1163 (2010)

[2] Wibisono, A., Wilson, A.C., Jordan, M.I.: A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences pp. 201614,734 0027–8424 (2016)

[3] Bolte, J., Teboulle, M.: Barrier operators and associated gradient-like dynamical systems for constrained minimization problems. SIAM Journal on Control and Optimization **42**(4), 1266–1292 (2003)

[4] Alvarez, F., Bolte, J., Brahic, O.: Hessian Riemannian gradient flows in convex programming. SIAM Journal on Control and Optimization **43**(2), 477–501 (2004)

[5] Attouch, H., Bolte, J., Redont, P., Teboulle, M.: Singular Riemannian barrier methods and gradient-projection dynamical systems for constrained optimization. Optimization **53**(5-6), 435–454 (2004)

[6] Nesterov, Y.: Dual extrapolation and its applications to solving variational inequalities and related problems. Mathematical Programming **109**(2), 319–344 (2007). DOI 10.1007/s10107-006-0034-z. URL http://dx.doi.org/10.1007/s10107-006-0034-z

[7] Nesterov, Y.: Primal-dual subgradient methods for convex problems. Mathematical Programming **120**(1), 221–259 (2009)

[8] Nemirovski, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York, NY (1983)

[9] Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. Journal of Machine Learning Research **11**(Oct), 2543–2596 (2010)

[10] Duchi, J.C., Agarwal, A., Wainwright, M.J.: Dual averaging for distributed optimization: Convergence analysis and network scaling. IEEE Transactions on Automatic control **57**(3), 592–606 (2012)

[11] Mertikopoulos, P., Staudigl, M.: On the convergence of gradient-like flows with noisy gradient input. forthcoming: Siam Journal of Optimization (2017)

[12] Facchinei, F., Pang, J.S.: Finite-dimensional variational inequalities and complementarity problems. Springer (2003)

[13] Harker, P.T., Pang, J.S.: Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications. Mathematical Programming **48**(1), 161–220 (1990). DOI 10.1007/BF01582255. URL http://dx.doi.org/10.1007/BF01582255

[14] Nemirovski, A., Onn, S., Rothblum, U.G.: Accuracy certificates for computational problems with convex structure. Mathematics of Operations Research **35**(1), 52–78 (2010)

[15] Ferris, M.C., Pang, J.S.: Engineering and economic applications of complementarity problems. Siam Review **39**(4), 669–713 (1997)

[16] Scutari, G., Palomar, D.P., Facchinei, F., Pang, J.s.: Convex optimization, game theory, and variational inequality theory. IEEE Signal Processing Magazine **27**(3), 35–49 (2010)

[17] Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, *A Series of Comprehensive Studies in Mathematics*, vol. 317. Springer-Verlag, Berlin (1998)

[18] Mertikopoulos, P.: Learning in games with continuous action sets and unknown payoff functions. https://arxiv.org/abs/1608.07310 (2016)

[19] Shalev-Shwartz, S.: Online learning and online convex optimization. Foundations and Trends in Machine Learning **4**(2), 107–194 (2011)

[20] Zălinescu, C.: Convex analysis in general vector spaces. World Scientific Publishing Co. (2002)

[21] Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton, NJ (1970)

[22] Borwein, J.M., Dutta, J.: Maximal monotone inclusions and fitzpatrick functions. Journal of Optimization Theory and Applications **171**(3), 757–784 (2016). DOI 10.1007/s10957-015-0813-x. URL http://dx.doi.org/10.1007/s10957-015-0813-x

[23] Nikaido, H., Isoda, K.: Note on non-cooperative convex games. Pacific Journal of Mathematics **5**, 807–815 (1955). URL https://projecteuclid.org:443/euclid.pjm/1171984836

[24] Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics **7**(3), 200–217 (1967)

[25] Kiwiel, K.C.: Free-steering relaxation methods for problems with strictly convex costs and linear constraints. Mathematics of Operations Research **22**(2), 326–349 (1997)

[26] Bruck Jr., R.E.: On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space. Journal of Mathematical Analysis and Applications **61**(1), 159–164 (1977). DOI http://dx.doi.org/10.1016/0022-247X(77)90152-4. URL //www.sciencedirect.com/science/article/pii/0022247X77901524

[27] Nemirovski, A.: Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM Journal on Optimization **15**(1), 229–251 (2004). DOI 10.1137/S1052623403425629. URL http://dx.doi.org/10.1137/S1052623403425629

[28] Kwon, J., Mertikopoulos, P.: A continuous-time approach to online optimization. Journal of Dynamics and Games **4**(2), 125–148 (2017)

[29] Khasminskii, R.Z.: Stochastic Stability of Differential Equations, 2 edn. No. 66 in Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin (2012)

[30] Benaïm, M., Hirsch, M.W.: Asymptotic pseudotrajectories and chain recurrent flows, with applications. Journal of Dynamics and Differential Equations **8**(1), 141–176 (1996)

[31] Karatzas, I., Shreve, S.E.: Brownian Motion and Stochastic Calculus. Springer-Verlag, Berlin (1998)

[32] Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. No. 87 in Applied Optimization. Kluwer Academic Publishers (2004)

[33] Yong, J., Zhou, X.Y.: Stochastic Controls- Hamiltonian Systems and HJB equations. Springer (1999)