STOCHASTIC MONOTONE INCLUSION WITH CLOSED LOOP DISTRIBUTIONS

HAMZA ENNAJI[c,*], JALAL FADILI[♯], AND HEDY ATTOUCH[◇]

Dedicated to the memory of Hedy Attouch, outstanding mathematician and beloved collaborator.

Abstract. In this paper, we study in a Hilbertian setting, first and second-order monotone inclusions related to stochastic optimization problems with decision-dependent distributions. The studied dynamics are formulated as monotone inclusions governed by Lipschitz perturbations of maximally monotone operators where the concept of equilibrium plays a central role. We discuss the relationship between the W_1-Wasserstein Lipschitz behavior of the distribution and the so-called coarse Ricci curvature. As an application, we consider the monotone inclusions associated with stochastic optimisation problems involving the sum of a smooth function with Lipschitz gradient, a proximable function and a composite term.

1. Introduction

Recently, many problems in machine learning and risk management come in form of stochastic optimisation problems. Such problems aim to learn a decision rule from a data sample. This can be formulated in terms of optimization problems of the form

$$\min_{x\in\mathbb{R}^n} \mathbb{E}_{\xi\sim\mathsf{m}}(f(x,\xi)) + g(x), \tag{1}$$

where m is a probability measure, $\mathbb{E}_{\xi\sim\mathsf{m}}$ is the expectation operator with respect to the measure m, $f(x,\xi)$ is a loss function of the decision x at data point ξ and g is a regularizer. In this work, we are interested in (1) in the case where the distribution m depends itself on the decision x, i.e., problems of the form

$$\min_{x\in\mathbb{R}^n} \mathbb{E}_{\xi\sim\mathsf{m}_x}(f(x,\xi)) + g(x), \tag{2}$$

In this case, one tries to learn a decision rule from a decision-dependent data distribution m_x. Problems of the form (2) were addressed in the framework of performative prediction proposed in [42, 46] and discussed with further algorithmic aspects in [34]. A typical example concerns prediction of loan default risks, that is the chance that a borrower won't be able to repay their loan. More precisely, banks take into account several parameters, including the default risk, to decide whether to accept a consumer's loan application and, if so, what interest rate will apply. It is clear that a high default risk implies a high interest rate, but a high interest rate increases the consumer's default risk. Thus, the predictive performance of the bank's model is not calibrated with respect to future results obtained by acting on

[c] Corresponding author.

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

[♯] ENSICAEN, Normandie Université, CNRS, GREYC, France.

[◇] IMAG, Université Montpellier, CNRS, France.

E-mail addresses: hamza.ennaji@univ-grenoble-alpes.fr, jalal.fadili@ensicaen.fr, hedy.attouch@umontpellier.fr.

2020 *Mathematics Subject Classification.* Primary 34G25, 37N40, 46N10, 47H05, 49M30, 60J20.

Key words and phrases. Monotone inclusions; Dynamical systems; Gradient flows; Inertial dynamics; Ollivier-Ricci curvature.

the model. Another example concerns navigation apps, such as Google Maps (see, e.g., [32, 40]), which suggest routes with low travel time to users. This influences users' decisions to pick such routes and consequently, increases traffic on these routes, impacting travel times. Further applications and illustrations can be found in [46, Appendix A].

In general, problems of the form (2) are difficult to solve. However, a natural approach consists in performing a repeated minimization procedure, i.e., throughout iterations, one solves

$$x_{t+1} \in \operatorname*{argmin}_{x\in\mathbb{R}^n} \mathbb{E}_{\xi\sim \mathsf{m}_{x_t}}(f(x,\xi)) + g(x), \tag{3}$$

and then updates the distribution $\mathsf{m}_{x_{t+1}}$. Under suitable assumptions that will be specified later, the sequence $(x_t)_t$ generated by the repeated minimization procedure (3) admits a fixed point $\bar{x}$. Such a point turns out to be an equilibrium with respect to the distribution $\mathsf{m}_{(.)}$ in the following sense:

$$\bar{x} \in \operatorname*{argmin}_{x\in\mathbb{R}^n} \ \mathbb{E}_{\xi\sim \mathsf{m}_{\bar{x}}}(f(x,\xi)) + g(x), \tag{4}$$

that is, $\bar{x}$ solves (2) for the induced distribution $\mathsf{m}_{\bar{x}}$. So instead of solving (2), we look at an equilibrium point in the sense of (4). Notice that in terms of operators, (4) can be written (formally, for instance) as the monotone inclusion:

$$A(\bar{x}) + B(\bar{x}) \ni 0 \tag{5}$$

with $A(x) = \partial g(x)$ and $B(x) = \mathbb{E}_{\xi\sim \mathsf{m}_{\bar{x}}}(\nabla f(x,\xi))$ (∇ being always the gradient w.r.t. to the variable x), where $A+B$ is monotone. That is $\bar{x}$ is a zero of the sum of two monotone operators. One strategy to solve problems of the form (5) is to consider some continuous and discrete dynamical systems whose trajectories may converge, under suitable assumptions, to an element in $(A+B)^{-1}(0)$, the zero set of the sum $A+B$. For instance, when $B\equiv 0$ and $A=\partial g$ where g is a proper convex lower semicontinuous function on $\mathbb{R}^n$, it is well known since the works of Brézis, Baillon and Bucker [21, 28], that each trajectory of the subgradient flow

$$\dot{x}(t) + \partial g(x) \ni 0 \text{ with } x(0) = x_0 \in \mathbb{R}^n, \tag{6}$$

converges to a minimizer of g, and thus a zero of ∂g, provided $\operatorname{argmin} g \neq \emptyset$.

Designing algorithms and dynamical systems with rapid convergence properties to solve monotone inclusions is at the core of many fields in modern optimization, partial differential equation, game theory, etc. The literature is extensive, and to name only a few, the reader is referred to [1, 2, 7, 9, 17, 25, 37, 39, 50] and the references therein.

In this work, we address closed-loop differential inclusions of the form

$$\dot{x}(t) + A(x(t)) + \mathbb{E}_{\xi\sim \mathsf{m}_{x(t)}}(B(x(t),\xi)) \ni 0, \tag{7}$$

where A is a maximally monotone operator, and $B(\cdot,\xi)$ is a single-valued mapping for all $\xi\in\Xi$. The particularity of such a dynamic is, of course, the presence of the random operator $\mathbb{E}_{\xi\sim \mathsf{m}_{(.)}}(B(.,\xi))$ where the random variable ξ has a trajectory-dependent distribution $\mathsf{m}_{x(t)}$. Thus, it is not straightforward how to address (7) within the classical framework (see, e.g., [22, 26]). Yet, a clever reformulation of (7), based on the notion of equilibrium, as a monotone inclusion governed by a Lipschitz perturbation of a maximally monotone operator will allow us to tackle this issue. Then we investigate inertial dynamics related to (7). Indeed, since the work of Polyak [47], who considered a system of the form

$$\ddot{x}(t) + \gamma\dot{x}(t) + \nabla f(x(t)) = 0, \tag{HBF}$$

where $\gamma > 0$ is called the viscous damping coefficient, the introduction of inertial dynamics to accelerate optimization methods has gained a lot of attention and led to many developments (see, e.g., [3, 10, 15, 18, 19, 50] and the references therein).

In this paper, we then consider second-order dynamics of the form

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f_{\mathsf{m}_{\bar{x}}}(x(t)) + \omega\nabla^2 f_{\mathsf{m}_{\bar{x}}}(x(t))\dot{x}(t) + \mathbf{e}_{\bar{x}}(x(t)) + \omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t)) = 0, \tag{8}$$

where $f_{\mathsf{m}_{\bar{x}}}(x) = \mathbb{E}_{\xi\sim\mathsf{m}_{\bar{x}}}(f(x,\xi))$, $\mathbf{e}_{\bar{x}} : \mathcal{H} \to \mathcal{H}$ is a perturbation operator and ω is the so-called Hessian-driven damping coefficient. When $\gamma(t) \equiv \gamma$, $f_{\mathsf{m}_{\bar{x}}} = f$ (i.e., without a stochastic structure) and $\mathbf{e}_{\bar{x}}(x) = 0$, systems of the form (8) were first studied in [3]. Later, this system was combined with an asymptotic vanishing damping $\gamma(t) = \frac{\alpha}{t}$, for $\alpha > 0$ in [19]. Several recent studies have been devoted to this topic (see, e.g., [9, 25, 37, 39, 49]).

1.1. Statement of the problem.

Throughout, $\mathcal{H}$ is a real Hilbert space endowed with the scalar product $\langle .,. \rangle$ and induced norm $\|\cdot\|$, and Ξ is a Polish space, i.e., separable and completely metrizable.

We consider the closed-loop differential inclusion

$$\begin{cases} \dot{x}(t) + A(x(t)) + \mathbb{E}_{\xi\sim\mathsf{m}_{x(t)}}\left(B(x(t),\xi)\right) \ni 0, \ \text{a.e } t > t_0 > 0, \\ x(t_0) = x_0 \in \overline{\mathrm{dom}(A)}, \end{cases} \tag{9}$$

where m_x is a family of probability distributions on Ξ indexed by $x \in \mathcal{H}$. We will work under the standing assumption:

Assumption 1.

- $A : \mathcal{H} \rightrightarrows \mathcal{H}$ is a set-valued maximal monotone operator such that $\mathrm{int}(\mathrm{dom}(A)) \neq \emptyset$;
- $B : \mathcal{H} \times \Xi \to \mathcal{H}$ is single-valued with $\xi \mapsto B(x,\xi)$ m_x-measurable, and $\exists\beta > 0$ such that $\xi \mapsto B(x,\xi)$ is β-Lipschitz continuous for every $x \in \mathcal{H}$;

Observe that when $x \mapsto \mathsf{m}_x(C)$ is a measurable function on $\mathcal{H}$ for each fixed $C \in \mathcal{B}$, where $\mathcal{B}$ is a countably generated σ-field on Ξ, $(\mathsf{m}_x)_{x\in\mathcal{H}}$ can be viewed as a random walk on $(\mathcal{H} \times \Xi, \mathcal{F} \otimes \mathcal{B})$ where $\mathcal{F}$ is a σ-field on $\mathcal{H}$; see Section 5.

► **Example 1.1.**

1. Typically, (9) covers the case of stochastic optimization problems with a state-dependent distribution studied in [34, 46] by taking $A = \partial g$ and $B = \nabla f$ for $g \in \Gamma_0(\mathcal{H})$ and $f \in C^1(\mathcal{H} \times \xi)$ whose gradient $\nabla f(x,\cdot)$ is β-Lipschitz continuous for every $x \in \mathcal{H}$. The last assumption ensures that $\mathbb{E}_{\xi\sim\mathsf{m}}\left(f(\cdot,\xi)\right)$ is $C^1(\mathcal{H})$ whose gradient is $\mathbb{E}_{\xi\sim\mathsf{m}}\left(\nabla f(\cdot,\xi)\right)$ which is β-Lipschitz continuous.
2. Taking $A = N_K$ the normal cone of a nonempty closed convex set K of admissible decisions, we recover the framework of variational inequalities addressed recently in [31].

To simplify the presentation, we set, for any measure $\mathsf{m} \in \mathcal{P}(\Xi)$

$$B_{\mathsf{m}}(x) = \mathbb{E}_{\xi\sim\mathsf{m}}\left(B(x,\xi)\right) \ \text{ and } \ F_{\mathsf{m}}(x) = A(x) + B_{\mathsf{m}}(x). \tag{10}$$

Using this notation, the system (9) can be simply rewritten as

$$\begin{cases} \dot{x}(t) + F_{\mathsf{m}_{x(t)}}(x(t)) \ni 0, \ \text{a.e } t > t_0, \\ x(t_0) = x_0 \in \overline{\mathrm{dom}(A)}. \end{cases} \tag{SMI}$$

The acronym (SMI) for the above dynamic stands for Stochastic Monotone Inclusion. Though "monotone" may seem as an abuse of terminology because the measure $\mathsf{m}_{(.)}$ depends on the trajectory, and B_{m} is not even monotone. We will show later that this terminilogy is still justified as (SMI) can be reformulated as a Lipschitzian perturbation of a monotone inclusion (see Section 3.2).

1.2. **Contributions and organization of the paper.** The paper is organized as follows. In Section 3 we address first-order monotone inclusions with closed-loop distributions. We prove the existence of equilibria as well as the well-posedness of the dynamics and convergence properties of the trajectories. In Section 4 we study asymptotic convergence properties of the trajectories of second-order dynamics with closed-loop distributions via Hessian damping. This allows us in particular to cover problems of the form (2). Section 5 contains a discussion concerning the Lipschitz behavior of the family $(\mathsf{m}_x)_{x\in\mathcal{H}}$ with respect to the Wasserstein distance and some consequences in the framework of Markov chains on metric random walk spaces. In Section 6 we discuss the inertial primal-dual algorithm as an application of our results. Finally, Section 7 contains some conclusions and discusses some future works.

2. Notation and preliminaries

In this section we fix some notation and present some notions and results that will be used.

2.1. **Convex analysis.** The domain of a function g on $\mathcal{H}$ is defined by $\mathrm{dom}(g) = \{x \in \mathcal{H} :\ g(x) < \infty\}$. We denote by $\Gamma_0(\mathcal{H})$ the class of proper (bounded from below and $\mathrm{dom}(g) \neq \emptyset$), lower semicontinuous (l.s.c) and convex functions on $\mathcal{H}$ with values in $\mathbb{R}\cup\{+\infty\}$. We say that g is α-strongly convex, for $\alpha > 0$, if $g - \frac{\alpha}{2}\|.\|^2$ is convex.

The subdifferential of g is defined as

$$\partial g : x \in \mathcal{H} \mapsto \{v \in \mathcal{H} :\ g(y) \geq g(x) + \langle v, y - x\rangle\}.$$

We recall the following Fermat's optimality condition for $g \in \Gamma_0(\mathcal{H})$,

$$0 \in \partial g(x^\star) \Leftrightarrow x^\star \in \operatorname{argmin} g(\mathcal{H}).$$

Definition 1 (Differentiability)**.** Let $g : \mathcal{H} \to \mathbb{R}\cup\{\infty\}$ and $x \in \mathrm{int}(\mathrm{dom}(g))$. We say that g is (Fréchet) differentiable at x if there exists $v \in \mathcal{H}$ such that

$$\lim_{h\to 0} \frac{g(x+h) - g(x) - \langle v, h\rangle}{\|h\|} = 0.$$

The unique vector v satisfying this condition is the gradient of g at x denoted by $\nabla g(x)$.

If $g \in \Gamma_0(\mathcal{H})$ and differentiable at x, then $\partial g(x) = \{\nabla g(x)\}$.

We also recall the following.

Definition 2 (L-smoothness)**.** Let $L \geq 0$ and $g : \mathcal{H} \to \mathbb{R}\cup\{\infty\}$. We say that g is L-smooth over $D \subset \mathcal{H}$ if it is differentiable over D and

$$\|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\| \text{ for any } x, y \in D.$$

We denote by $C_L^{1,1}(D)$ the class of L-smooth functions over D.

Given a function $g : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$, it proximal mapping is defined through

$$\mathrm{prox}_f(x) = \operatorname*{argmin}_{y\in\mathcal{H}} \left\{ g(y) + \frac{1}{2}\|y - x\|^2 \right\} \text{ for any } x \in \mathcal{H}.$$

When $g = \iota_K$ the indicator function of a nonempty closed convex set $K \subset \mathcal{H}$, then $\mathrm{prox}_f = \mathrm{proj}_K$, the projector onto K. For further details and notion, we refer the reader to [23].

2.2. **Operator theory.** The domain of the set-valued operator $A:\mathcal{H}\rightrightarrows\mathcal{H}$ is $\operatorname{dom}(A)=\{x\in\mathcal{H}:\ A(x)\neq\emptyset\}$, its graph is $\operatorname{gra}(A)=\{[x,u]\in\mathcal{H}\times\mathcal{H}:\ u\in Ax\}$ and its zeros set is $\operatorname{zer}(A)=\{x\in\mathcal{H}:\ 0\in A(x)\}:=A^{-1}(0)$. A selection of A is an operator $T:\operatorname{dom}A\to\mathcal{H}$ such that, $Tx\in Ax$ for any $x\in\operatorname{dom}A$. We write $A:\mathcal{H}\to\mathcal{H}$ to indicate that A is single-valued. In the following, we gather some main properties that are essential for the rest of the paper.

Definition 3.

- We say that A is β-Lipschitz continuous if it is single-valued over its domain and
$$\|Ax-Ay\|\leq\beta\|x-y\|,\quad\forall x,y\in\operatorname{dom}A.\tag{11}$$
- We say that $A:\mathcal{H}\rightrightarrows\mathcal{H}$ is monotone if
$$\langle x-y,u-v\rangle\geq 0,\quad\forall[x,u],[y,v]\in\text{gra }A.\tag{12}$$
- We say that A is maximal monotone if there exists no monotone operator B, i.e., satisfying (12), such that gra $A\subset$ gra B.
- We say that A is uniformly monotone with modulus $\phi:[0,\infty)\to[0,\infty)$ if ϕ is increasing, $\phi(0)=0$, $\lim_{t\to\infty}\phi(t)=\infty$ and
$$\langle x-y,u-v\rangle\geq\|x-y\|\phi\left(\|x-y\|\right),\quad\forall[x,u],[y,v]\in\text{gra }A.\tag{13}$$
- We say that A is μ-strongly monotone, with $\mu>0$, if
$$\langle x-y,u-v\rangle\geq\mu\|x-y\|^2,\quad\forall[x,u],[y,v]\in\text{gra }A.\tag{14}$$

Remark 2.1.

- Note that if A is μ-strongly montone is equivalent to saying that $A-\mu\mathrm{Id}$ is monotone.
- The definition of uniform monotonicity given by (13) is slightly different from the one in [23, Definition 22.1].
- If A is μ-strongly monotone, then it is uniformly monotone with modulus $\phi(t)=\mu t$.

► **Example 2.1.**

- The typical example of a maximal monotone operator is the subdifferential ∂g of a function $g\in\Gamma_0(\mathcal{H})$. We usually refer to such an operator as a subpotential maximal monotone operator.

2.3. **Monotone inclusions.** Let A be a maximal monotone operator on $\mathcal{H}$ and a single-valued mapping $D:[t_0,+\infty[\times\overline{\operatorname{dom}(A)}\to\mathcal{H}$ and consider the following differential inclusion
$$\begin{cases}\dot{x}(t)+A(x(t))+D(t,x(t))\ni 0, & t\in[t_0,T]\\ x(t_0)=x_0\in\operatorname{dom}(A).\end{cases}\tag{15}$$

Definition 4. We will say that $x:[t_0,T]\to\mathcal{H}$ is a strong solution trajectory on $[t_0,T]$ of (15) if the following properties are satisfied:

(a) x is continuous on $[t_0,T]$ and absolutely continuous on any compact subset of $]t_0,T[$ (hence almost everywhere differentiable);

(b) $x(t)\in\operatorname{dom}(A)$ for almost every $t\in]t_0,T]$, and (15) is verified for almost every $t\in]t_0,T[$.

A trajectory $x:[t_0,+\infty[\to\mathcal{H}$ is a strong global solution of (15) if it is a strong solution on $[t_0,T]$ for any $T>t_0$.

For further details, we refer the reader to the classical monographs [26] or [22].

2.4. **Transportation distance.** Denote $\mathcal{P}(\Xi)$ the space of probability measures on Ξ. For $\mathsf{m}_1, \mathsf{m}_2 \in \mathcal{P}(\Xi)$, the $\mathbb{W}_1$-Wasserstein distance is defined by

$$\mathbb{W}_1(\mathsf{m}_1, \mathsf{m}_2) = \sup_{h \in \mathrm{Lip}_1} \left| \mathbb{E}_{\xi \sim \mathsf{m}_1} h(\xi) - \mathbb{E}_{\zeta \sim \mathsf{m}_2} h(\zeta) \right|, \tag{16}$$

where Lip_1 is the space of 1-Lipschitz continuous functions $h : \Xi \to \mathbb{R}$.

3. First-order monotone inclusions

In this section we perform the analysis of the first-order monotone inclusion (9). More precisely, we discuss the existence and uniqueness of solutions as well as the convergence of trajectories. Recall that the dynamic (9) is governed by the operator $F_{\mathsf{m}_x} = A + B_{\mathsf{m}_x}$. We first prove the existence of an equilibrium point $\bar{x}$ which will allow us to reformulate (9) in a suitable form.

The following assumption is essential for the convergence analysis. It describes the sensitivity of the distribution to shifts in the index (here trajectory). It is widely used in the literature (see, e.g., [34, 42, 46, 52]). We will discuss how it closely relates to the so-called coarse Ricci curvature in Section 5.

Assumption 2 (Lipschitz distributions)**.** There exists $\tau > 0$ such that

$$\mathbb{W}_1(\mathsf{m}_x, \mathsf{m}_y) \leq \tau \|x - y\|, \text{ for all } x, y \in \mathcal{H}.$$

The following two assumptions are standard monotonicity assumptions and will be crucial for the well-posedness of the dynamics and to prove existence and uniqueness of equilibria.

Assumption 3 (Strong monotonicity)**.** $\exists \mu > 0$ such that for all $x \in \mathcal{H}$, F_{m_x} is μ-strongly monotone.

The strong monotonicity Assumption 3 can be weakened to uniform monotonicity in the following sense.

Assumption 4 (Uniform monotonicity)**.** There exists a function ϕ satisfying

$$\phi(t) > \beta\tau t, \ \forall t > 0,$$

such that for all $x \in \mathcal{H}$, F_{m_x} is uniformly monotone with a modulus ϕ.

Finally, define the following parameter $\rho := \frac{\beta\tau}{\mu}$. As we will see, (see also [34, 46]), the parameter regime $\rho < 1$ will play a crucial role in the analysis of the convergence of the trajectories.

3.1. Existence and uniqueness of equilibria.

Definition 5. (Equilibrium point) We say that $\bar{x} \in \mathcal{H}$ is at equilibrium with respect to the family of probability measures $(\mathsf{m}_x)_{x \in \mathcal{H}}$ if

$$0 \in \ F_{\mathsf{m}_{\bar{x}}}(\bar{x}). \tag{17}$$

In case $A = \partial g$ and $B = \nabla f$ where $g \in \Gamma_0(\mathcal{H})$ and $f(x, \cdot) \in C^{1,1}_\beta(\Xi)$, this definition is to be compared to the one introduced in [46] (see also [34]). Indeed, (17) reduces to:

$$\bar{x} \in \operatorname*{argmin}_{x \in \mathcal{H}} \mathbb{E}_{\xi \sim \mathsf{m}_{\bar{x}}}(f(x, \xi)) + g(x). \tag{18}$$

Solutions of (18) are exactly the fixed points of the repeated minimization procedure, that is, starting from some x_0, we generate the following sequence for $t \geq 0$

$$x_{t+1} = S(x_t) := \operatorname*{argmin}_{x \in \mathcal{H}} \mathbb{E}_{\xi \sim \mathsf{m}_{x_t}}(f(x, \xi)) + g(x). \tag{19}$$

In [46, Theorem 3.5] it is shown that if f is C^1 in both variables, $\xi \mapsto \nabla f(x,\xi)$ is $\beta-$Lipschitz and $\mathbb{E}_{\xi\sim \mathsf{m}_x}(f(.,\xi))$ is μ-strongly convex for all $x \in \mathcal{H}$ with $\rho < 1$, then, under Assumption 2, the iterates of (19) converge to a unique stable point. Their proof is essentially based on a fixed point argument. In [46, Propostion 4.1], they show the existence of equilibrium points under weaker assumptions on the loss f. Specifically, they demonstrate that if f is convex and jointly continuous, then equilibrium points exist provided $\mathrm{dom}(g)$ is compact. However, in this case the equilibrium is not necessarily unique. In the following lemma, we show the existence of equilibrium in the sense of (17).

Theorem 1 (Existence and uniqueness of equilibrium point)**.** *Under Assumption 4, the map*

$$S : x \in \mathcal{H} \mapsto \mathrm{zer}(F_{\mathsf{m}_x}) = \{u \in \mathcal{H} :\ 0 \in F_{\mathsf{m}_x}(u)\},$$

is a contraction. In particular, the equilibrium $\bar{x}$ is unique. If moreover, Assumption 3 holds instead, i.e., F_{m_x} is μ-strongly monotone for $\mu > 0$, the mapping S is ρ-Lipschitz with $\rho := \frac{\beta\tau}{\mu}$. Thus for $\rho < 1$, there exists a unique equilibrium point $\bar{x}$.

Proof. First, we see that S is well defined. Indeed, for any $x \in \mathcal{H}$, $\mathrm{zer}(F_{\mathsf{m}_x})$ is nonempty, and is in fact a singleton due to Assumption 4. To see this, we argue as in [23, Proposition 22.11]. Fix $[y_0,u_0] \in \mathrm{gra}\ F_{\mathsf{m}_x}$. We have for any $[y,u] \in \mathrm{gra}\ F_{\mathsf{m}_x}$:

$$\begin{aligned}\|y-y_0\|\|u\| &\geq \langle y-y_0,u\rangle = \langle y-y_0,u-u_0\rangle + \langle y-y_0,u_0\rangle \\ &\geq \|y-y_0\|\phi\left(\|y-y_0\|\right) - \|y-y_0\|\|u_0\|.\end{aligned} \tag{20}$$

Since $\lim_{t\to\infty}\phi(t) = \infty$ we infer that $\inf_{u\in \mathrm{gra}\ F_{\mathsf{m}_x}(y)}\|u\|\to\infty$ as $\|y\|\to\infty$ and thus F_{m_x} is surjective (see [23, Corollary 21.25]). Moreover, in view of strict monotonicity of F_{m_x}, $\mathrm{zer}(F_{\mathsf{m}_x})$ is a singleton (see, e.g., [23, Proposition 23.35]).

Now, pick $x,y \in \mathcal{H}$. We have that $0 \in F_{\mathsf{m}_x}(S(x))$ and $0 \in F_{\mathsf{m}_y}(S(y))$. In particular, $-B_{\mathsf{m}_y}(S(y)) \in A(S(y))$, which gives that $B_{\mathsf{m}_x}(S(y)) - B_{\mathsf{m}_y}(S(y)) \in F_{\mathsf{m}_x}(S(y))$. We get, thanks to Assumption 4

$$\|S(x)-S(y)\|\phi\Big(\|S(x)-S(y)\|\Big) \leq \langle u-v, S(x)-S(y)\rangle,\ \text{for any}\ (u,v) \in F_{\mathsf{m}_x}(S(y))\times F_{\mathsf{m}_x}(S(x)).$$

Then, taking $u = B_{\mathsf{m}_x}(S(y)) - B_{\mathsf{m}_y}(S(y))$ and $v = 0$, we get, using Cauchy-Schwarz inequality

$$\phi\Big(\|S(x)-S(y)\|\Big) \leq \|B_{\mathsf{m}_x}(S(y)) - B_{\mathsf{m}_y}(S(y))\|.$$

We get, using Corollary 1 below

$$\phi\Big(\|S(x)-S(y)\|\Big) \leq \beta\tau\|x-y\|. \tag{21}$$

Since ϕ is strictly increasing, the last inequality gives, thanks to Assumption 4,

$$\|S(x)-S(y)\| \leq \phi^{-1}\Big(\beta\tau\|x-y\|\Big) < \|x-y\|,$$

and by the Banach fixed-point theorem (see Theorem 6), S has a unique fixed point $\bar{x}$.

Now if F_{m_x} is μ-strongly monotone, it is in particular uniformly monotone with modulus $\phi(t) = \mu t$. Equation (21) gives

$$\|S(x)-S(y)\| \leq \rho\|x-y\|,$$

with $\frac{\beta\tau}{\mu} := \rho$. Again, we conclude using Theorem 6. ■

Remark 3.1. In the strongly monotone case, Assumption 4 incorporates the parameter regime $\rho < 1$ which appears in particular in [46, Theorem 3.5].

3.2. **Well-posedness.** Before stating the main result of this section, let us fix some extra notation and properties. Assume that Assumption 4 holds, and denote by

$$\varphi(t) = \phi(t) - \beta\tau t, \tag{22}$$

where ϕ is the modulus of uniform monotonicity of $F_{\mathsf{m}_{\bar{x}}}$.

We prove the following.

Lemma 1. *Let $a > 0$ and define*

$$\theta(z) := \int_z^a \frac{\mathrm{d}s}{\varphi(s)}. \tag{23}$$

Then θ is nonincreasing and $\lim_{z\to 0^+} \theta(z) = \infty$.

Proof. Indeed, we have $\dot{\theta}(z) = -1/\varphi(z) < 0$ since ϕ satisfies Assumption 4. Moreover, since $\phi(t) \geq \varphi(t)$, and $\lim_{z\to 0^+} \int_z^a \frac{\mathrm{d}s}{\phi(s)} = \infty$, the result follows. ■

Remark 3.2. In the literature of ordinary differential equations, the above lemma is related to the fact that φ is somehow an *Osgood modulus of continuity* (see, e.g., [20, Definitions 2.108 and 3.1]). If $\varphi(s) = s$, which corresponds to Lipschitz regularity, and $a = 1$ then $\theta(z) = \log_+(z) = \max\{0, \log(1/z)\}$. If $a = 1/e$ and $\varphi(s) = s\log(1/s)$, which corresponds to log-Lipschitz regularity, then $\theta(z) = \log\log_+(z)$. More generally, $\varphi(s) = s\left(\log(1/s)\right)^r$ for $r \leq 1$ are admissible choices.

Now let us define the following gap

$$\mathbf{e}_{\bar{x}}(x) = B_{\mathsf{m}_x}(x) - B_{\mathsf{m}_{\bar{x}}}(x). \tag{24}$$

Using the notation in (10), we may rewrite (9) in the following form

$$\begin{cases} \dot{x}(t) + F_{\mathsf{m}_{\bar{x}}}(x(t)) + \mathbf{e}_{\bar{x}}(x(t)) \ni 0, \text{ a.e } t > t_0 \\ x(t_0) = x_0. \end{cases} \tag{p-SMI}$$

One advantage of this formulation is that the mapping $x \mapsto \mathbf{e}_{\bar{x}}(x)$ exhibits Lipschitz behaviour (see Lemma 4), and now only the operator $F_{\mathsf{m}_{\bar{x}}}$ appears instead of $F_{\mathsf{m}_{x(t)}}$. This allows us to treat (p-SMI) within the framework of evolution equations governed by Lipschitz perturbations of maximal monotone operators (cf. [26, Chapter III]). This is behind the notatiton (p-SMI), which stands for perturbed stochastic monotone inclusion. This being said, our aim is to use [26, Proposition 3.13] and show the existence of a unique strong solution (cf. Definition 4) to (p-SMI) (see also [26, Definition 3.1]). To this end we begin with the following lemmas.

We start with the following properties.

Lemma 2. *Under Assumption 1, we have, for any $\mathsf{m}, \nu \in \mathcal{P}(\Xi)$ and $x \in \mathcal{H}$*

$$\sup_{x\in\mathcal{H}} \| B_{\mathsf{m}}(x) - B_\nu(x) \| \leq \beta \mathbb{W}_1(\mathsf{m}, \nu).$$

Proof. Let $v \in \mathcal{H}$ be a unit norm vector. By Assumption 1, $\langle v, B(x, \cdot)\rangle$ is β-Lipschitz continuous for every $x \in \mathcal{H}$. We then have from the definition of the $\mathbb{W}_1$-Wasserstein distance (16)

$$\begin{aligned} \| B_{\mathsf{m}}(x) - B_\nu(x) \| &= \sup_{v\in\mathcal{H}, \|v\|=1} \langle v, B_{\mathsf{m}}(x) - B_\nu(x) \rangle \\ &= \sup_{v\in\mathcal{H}, \|v\|=1} \mathbb{E}_{\xi\sim\mathsf{m}} \langle v, B(x,\xi)\rangle - \mathbb{E}_{\zeta\sim\nu}\langle v, B(x,\zeta)\rangle \leq \beta\mathbb{W}_1(\mathsf{m}, \nu). \end{aligned}$$

Taking the supremum over $x \in \mathcal{H}$ yields the result. ■

Combining Lemma 2 and Assumption 2 we get the following.

Corollary 1. *Under Assumption 1 and Assumption 2, for all $y, z \in \mathcal{H}$*

$$\sup_{x\in\mathcal{H}} \|B_{\mathsf{m}_y}(x) - B_{\mathsf{m}_z}(x)\| \leq \beta\tau\|y - z\|.$$

Assumption 5 (Lipschitz continuity of B_{m_x})**.** $\exists L > 0$ such that B_{m_x} is L-Lipschitz continuous for every $x \in \mathcal{H}$.

This assumption is true if for instance $B(\cdot, \xi)$ is L-Lipschitz continuous for m_x-almost every $\xi \in \Xi$. Indeed

$$B_{\mathsf{m}_x}(z) - B_{\mathsf{m}_x}(y) = \mathbb{E}_{\xi\sim\mathsf{m}_x}(B(z, \xi) - B(y, \xi)) \leq L\|z - y\|.$$

We are now in position to characterize the properties of $F_{\mathsf{m}_{\bar{x}}}$.

Lemma 3. *Assume that Assumptions 1 and 5 hold and that $B_{\mathsf{m}_{\bar{x}}}$ is monotone. Then the operator $F_{\mathsf{m}_{\bar{x}}}$ is maximally monotone. If, in addition, Assumption 3 (resp. Assumption 4) then $F_{\mathsf{m}_{\bar{x}}}$ is maximally uniformly (resp. strongly) monotone.*

Proof. Maximal monotonicity of $B_{\mathsf{m}_{\bar{x}}}$ follows from [23, Corollary 20.28]. Combine this with maximal monotonicity of A stated in Assumption 1, and [26, Lemma 2.4] yields the claim. The proof of the the last statement is immediate. ■

The key ingredient to use [26, Proposition 3.13] is the Lipschitz continuity of the perturbation $\mathbf{e}_{\bar{x}}(.)$. This is the content of the following statement.

Lemma 4. *Suppose that Assumptions 1, 2 and 5 hold, then $\mathbf{e}_{\bar{x}}(.)$ is $(2L + \beta\tau)$-Lipschitz continuous.*

Proof. Let $x, z \in \mathcal{H}$. We then have

$$\begin{aligned}
\|\mathbf{e}_{\bar{x}}(x) - \mathbf{e}_{\bar{x}}(z)\| &= \|(B_{\mathsf{m}_x}(x) - B_{\mathsf{m}_{\bar{x}}}(x)) - (B_{\mathsf{m}_z}(z) - B_{\mathsf{m}_{\bar{x}}}(z))\,\| \\
&= \|(B_{\mathsf{m}_x}(x) - B_{\mathsf{m}_x}(z)) + (B_{\mathsf{m}_{\bar{x}}}(z) - B_{\mathsf{m}_{\bar{x}}}(x)) + (B_{\mathsf{m}_x}(z) - B_{\mathsf{m}_z}(z))\,\| \\
&\leq L\|x - z\| + L\|x - z\| + \beta\tau\|x - z\| \\
&= (2L + \beta\tau)\|x - z\|,
\end{aligned} \tag{25}$$

where we have used Assumption 5 twice and Corollary 1 once in the inequality. ■

Remark 3.3. Since $\operatorname{zer}(F_{\mathsf{m}_{\bar{x}}}) = \{\bar{x}\}$, we clearly see that $\bar{x} \in \operatorname{zer}(F_{\mathsf{m}_{\bar{x}}} + \mathbf{e}_{\bar{x}})$. Indeed, taking into account (24), we have $\mathbf{e}_{\bar{x}}(\bar{x}) = 0$ so that $0 \in (F_{\mathsf{m}_{\bar{x}}} + \mathbf{e}_{\bar{x}})(\bar{x})$.

Proposition 1. *Suppose that Assumptions 1, 2 and 5 hold, that either Assumption 3 or Assumption 4 hold and let $T > t_0$. Then, given $x_0 \in \overline{\operatorname{dom}(F_{\mathsf{m}_{\bar{x}}})}$, the dynamic* (p-SMI)*, and hence* (SMI)*, admits a unique strong solution x in the sense of Definition 4.*

Proof. Thanks to Lemma 3 and Lemma 4 $F_{\mathsf{m}_{\bar{x}}}$ is maximally monotone and $\mathbf{e}_{\bar{x}}$ is Lipschitz continuous. Since $t \mapsto \mathbf{e}_{\bar{x}}(x)$ is trivially in $L^\infty(t_0, T; \mathcal{H})$ for any $x \in \mathcal{H}$, we deduce, thanks to [26, Proposition 3.13] the existence of a unique solution $x \in W^{1,1}(t_0, T; \mathcal{H})$ to (SMI). ■

Remark 3.4. Evolution problems of the form (p-SMI) were also studied in [13], where the operator $\mathbf{e}_{\bar{x}}$ is potentially multivalued and depends on both time and space variables. While standard time-variable measurability is assumed, the authors' analysis relied on the upper semicontinuity of $x \mapsto \mathbf{e}_{\bar{x}}(t, x)$, rather than Lipschitz continuity, for which weaker results were proved. We also observe that the domain condition in Assumption 1 can be removed in finite dimension (see e.g., [13, Theorem 1.2]) or when A is the subdifferential of a function in $\Gamma_0(\mathcal{H})$.

3.3. **Convergence properties.** We now establish the main convergence result of the unique solution trajectory of (SMI).

Theorem 2. *Let x be the unique solution trajectory of* (SMI) *under Assumptions 1, 2 and 5, and suppose that Assumption 4 also holds. Then, for all $t \geq t_0$, we have*

$$\|x(t) - \bar{x}\| \leq \theta^{-1}\Big(2t - \hat{t}\Big), \tag{26}$$

where θ is defined in Lemma 1 and $\hat{t} = 2t_0 - \theta\left(\|x_0 - \bar{x}\|\right)$.

Furthermore, if Assumption 3 holds instead and $\rho := \frac{\beta\tau}{\mu} < 1$, we have

$$\|x(t) - \bar{x}\| \leq Ce^{-2\mu(1-\rho)t} \text{ for all } t \geq t_0, \tag{27}$$

with $C = \|x_0 - \bar{x}\| e^{2\mu(1-\rho)t_0}$.

Proof. We observe thanks to Proposition 1 that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - \bar{x}\|^2 = \langle \dot{x}(t), x(t) - \bar{x}\rangle.$$

We then have, for any selection $u(t)$ of $F_{\mathsf{m}_{\bar{x}}}(x(t))$

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - y\|^2 &= \langle \dot{x}(t), x(t) - \bar{x}\rangle = -\langle u(t), x(t) - \bar{x}\rangle - \langle \mathbf{e}_{\bar{x}}(x(t)), x(t) - \bar{x}\rangle \\ &= \langle u(t) - 0, x(t) - \bar{x}\rangle - \langle \mathbf{e}_{\bar{x}}(x(t)), x(t) - \bar{x}\rangle. \end{aligned} \tag{28}$$

We then get by Assumption 4

$$\langle u(t) - 0, x(t) - \bar{x}\rangle \geq \|x(t) - \bar{x}\|\phi\left(\|x(t) - \bar{x}\|\right).$$

Then, thanks to Corollary 1, (28) gives

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - \bar{x}\|^2 &\leq -\|x(t) - \bar{x}\|\phi\left(\|x(t) - \bar{x}\|\right) + \|\mathbf{e}_{\bar{x}}(x(t))\|\|x(t) - \bar{x}\| \\ &\leq -\|x(t) - \bar{x}\|\phi\left(\|x(t) - \bar{x}\|\right) + \beta\tau\|x(t) - \bar{x}\|^2 \\ &= -\|x(t) - \bar{x}\|\varphi\left(\|x(t) - \bar{x}\|\right), \end{aligned} \tag{29}$$

where $\varphi(t) = \phi(t) - \beta\tau t$ is as introduced in (22). When $h(t) := \|x(t) - \bar{x}\| = 0$, then there is nothing to prove and we are done. Otherwise, $h(t) \neq 0$. Thus dividing both sides of (29) by $h(t)$, we get

$$\dot{h}(t) \leq -2\varphi\left(h(t)\right). \tag{30}$$

From (29) and (30), we infer that

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta\left(h(t)\right) = \frac{-1}{\varphi\left(h(t)\right)}\dot{h}(t) \geq 2. \tag{31}$$

Integrating (31) between t_0 and $t > 0$, we get

$$\theta(h(t)) - \theta(h(t_0)) \geq 2\left(t - t_0\right).$$

Using Lemma 1, we deduce that

$$h(t) \leq \theta^{-1}(2t - \hat{t}), \tag{32}$$

where $\hat{t} = 2t_0 - \theta(h(t_0))$. This proves (26), as desired.

To prove the second claim, we observe that if $F_{\mathsf{m}_{\bar{x}}}$ is μ-strongly monotone, $\phi(t) = \mu t$ so that $\varphi(t) = (\mu - \beta\tau)t$. Since $\mu > \beta\tau$, we get $\theta(s) = \frac{-\log(s)}{\mu - \beta\tau}$. So that $\theta^{-1}(s) = e^{-(\mu-\beta\tau)s}$. Plugging this into (32), we obtain

$$h(t) \leq \|x_0 - \bar{x}\|e^{-2(\mu-\beta\tau)(t-t_0)}, \tag{33}$$

as desired. ■

Remark 3.5. Observe that one can obtain (27) from (29) using Gronwall's Lemma 8. Indeed, taking $\phi(t) = \beta\mu t$, we infer from (29)

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)-\bar{x}\|^2 \leq -(\mu-\beta\tau)\|x(t)-\bar{x}\|^2.$$

By Assumption 2 and Theorem 2, we obtain a rate of convergence of the measure $\mathsf{m}_{x(t)}$ to $\mathsf{m}_{\bar{x}}$ in the $\mathbb{W}_1$ distance as $t \to \infty$.

Corollary 2. *Let $x : [t_0, \infty) \to \mathbb{R}$ be the solution of* (SMI) *under the assumptions of Proposition 1, and suppose that Assumption 4 also holds. We have for all $t \geq t_0$*

$$\mathbb{W}_1(\mathsf{m}_{x(t)}, \mathsf{m}_{\bar{x}}) \leq \tau\theta^{-1}\Big(2t - \hat{t}\Big),$$

where θ is defined in Lemma 1 and $\hat{t} = 2t_0 - \theta\left(\|x_0 - \bar{x}\|\right)$. If Assumption 3 holds instead and $\rho := \frac{\beta\tau}{\mu} < 1$, we have

$$\mathbb{W}_1(\mathsf{m}_{x(t)}, \mathsf{m}_{\bar{x}}) \leq Ce^{-2\mu(1-\rho)t},$$

with $C = \tau\|x_0 - \bar{x}\|e^{2\mu(1-\rho)t_0}$.

4. Inertial second-order system with viscous and Hessian damping

In this section, to avoid technicalities, we restrict ourselves to the smooth convex optimization case where i.e., $A = \nabla g$ so that (recall (10))

$$F_{\mathsf{m}} = \nabla G_{\mathsf{m}} \text{ with } G_{\mathsf{m}} := g + f_{\mathsf{m}} \text{ and } f_{\mathsf{m}} := \mathbb{E}_{\xi\sim\mathsf{m}}(f(\cdot,\xi)). \tag{34}$$

In this case, Assumptions 1 and 5 read as follows:

Assumption 6.

- $g \in C^1(\mathcal{H}) \cap \Gamma_0(\mathcal{H})$;
- $f \in C^1(\mathcal{H}\times\Xi)$, $f(x,\cdot)$ is m_x-measurable and $C^{1,1}_\beta(\Xi)$ for every $x \in \mathcal{H}$, and $f_{\mathsf{m}_x} \in C^{1,1}_L(\mathcal{H})$ for every $x \in \mathcal{H}$.

On the other hand, Assumption 3 specializes to

Assumption 7 (Strong convexity)**.** $\exists\mu > 0$ such that G_{m_x} is μ-strongly convex for every $x \in \mathcal{H}$.

This last assumption is true for instance if $f(\cdot,\xi)$ is μ-strongly convex for m_x-almost every $\xi \in \Xi$.

Let us first start with a discussion of the following second-order in time ODE

$$\begin{cases} \ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla G_{\mathsf{m}_{x(t)}}(x(t)) = 0, \text{ a.e } t \in [t_0, T] \\ x(t_0) = x_0 \in \mathcal{H}, \end{cases} \tag{35}$$

where $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ is a continuous function, usually referred to as the viscous damping coefficient. In a smooth convex optimization deterministic setting, i.e., $F_{\mathsf{m}_x} = \nabla f$ where f is a smooth, strongly convex function, systems of the form (35) were first studies by Polyak in [47] with a fixed viscous damping coefficient. Later on, this kind of systems were studied by [15] and then [50] where the authors establish the link between the continuous dynamics with $\gamma(t) = \frac{3}{t}$ and the Nesterov's gradient method [43]. This is a very active research field (see, e.g., [5, 11, 16, 38] and the references therein).

Following what we have done Section 3, (35) can be reformulated again as a Lipschitzian perturbation of an inertial gradient system as follows

$$\ddot{x}(t)+\gamma(t)\dot{x}(t)+\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))+\mathbf{e}_{\bar{x}}(x(t))=0. \tag{36}$$

Even in absence of perturbations, it is well-known known that the inertial system (36) may suffer from transverse oscillations, and it is desirable to attenuate them. This is precisely the motivation behind the introduction of geometric Hessian-driven damping in [3] (sometimes called Newton-type inertial dynamics). The inertial system we consider then features both viscous and Hessian-driven damping and reads

$$\ddot{x}(t)+\gamma(t)\dot{x}(t)+\omega(t)\frac{\mathrm{d}}{\mathrm{d}t}\left(\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\right)+\nabla G_{\mathsf{m}_{x(t)}}(x(t))=0,$$

where $\omega:[0,\infty)\to\mathbb{R}^+$ is a continuous function usually referred to as the Hessian-driven damping coefficient, and which will be taken to be constant, i.e., $\omega(t)\equiv\omega>0$. The case $\omega=0$ then recovers (35). Following the reasoning of Section 3, then under Assumptions 2, 6 and 7 this system is also equivalent to

$$\ddot{x}(t)+\gamma(t)\dot{x}(t)+\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))+\omega\frac{\mathrm{d}}{\mathrm{d}t}\left(\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\right)+\mathbf{e}_{\bar{x}}(x(t))+\omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t))=0 \text{ a.e } t\in[t_0,T]. \tag{ISEHD$_{\gamma,\omega}$}$$

Recall that $\mathbf{e}_{\bar{x}}$ is Lipschitz continuous by Lemma 4, and thus a.e. differentiable so that the last term in system (ISEHD$_{\gamma,\omega}$) makes sense whenever $x$is absolutely continuous.

The acronym ISEHD stands for Inertial System with Explicit Hessian Damping. The Hessian damping is said to be explicit since, when $G_{\mathsf{m}_{\bar{x}}}$ is of class C^2 and x is smooth, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\right)=\nabla^2 G_{\mathsf{m}_{\bar{x}}}(x(t))\dot{x}(t).$$

Variants and generalizations of the inertial systems with Hessian-driven damping were studied by multiple authors (see, e.g.,, [9, 18, 19]). The study of the effect of perturbations on these systems, i.e., problems of the form (ISEHD$_{\gamma,\omega}$), was carried out in the recent work [14]. We will take inspiration from this work but our analysis will depart from that of [14] in some important ways. For instance, and most importantly, the perturbations in [14] depend only on time while they depend on the trajectory in this paper. This poses a few challenges that we tackle by exploiting the Lipschitz continuity property enjoyed by our perturbations (see for instance Lemma 4).

4.1. **Well-posedness.**

4.1.1. *Equivalent first-order formulation.*

Proposition 2. *Suppose that Assumption 6 holds, that $\gamma(t)\geq 0,\omega>0$ with $\gamma\in C^1([t_0,+\infty[)$. For any initial conditions $(x_0,v_0)\in\mathcal{H}\times\mathcal{H}$, the dynamics (ISEHD$_{\gamma,\omega}$) admits an equivalent formulation of the form*

$$\begin{cases}\dot{x}(t)+\omega\Big(\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))+\mathbf{e}_{\bar{x}}(x(t))\Big)-\left(\frac{1}{\omega}-\gamma(t)\right)x(t)+\frac{1}{\omega}y(t) & =0\\ \dot{y}(t)-\left(\frac{1}{\omega}-\gamma(t)-\dot{\gamma}(t)\omega\right)x(t)+\frac{1}{\omega}y(t) & =0,\end{cases} \tag{37}$$

with initial conditions $x(t_0)=x_0, y(t_0)=-\omega\left(v_0+\omega\nabla G_{\mathsf{m}_{\bar{x}}}(x_0)\right)+(1-\omega\gamma(t_0))\,x_0-\omega^2\mathbf{e}_{\bar{x}}(x_0)$.

Proof. Let (x,y) be a solution of (37). By differentiation of the first equation in (37), we get

$$\ddot{x}(t)+\omega\frac{\mathrm{d}}{\mathrm{d}t}(\nabla G_{\mathsf{m}_{\bar{x}}}(x(t)))+\omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t))+\dot{\gamma}(t)x(t)-\Big(\frac{1}{\omega}-\gamma(t)\Big)\dot{x}(t)+\frac{1}{\omega}\dot{y}(t)=0.$$

Replacing $\dot{y}$ by its expression from the second equation in (37), we obtain:

$$\begin{aligned}\ddot{x}(t) + \omega\frac{\mathrm{d}}{\mathrm{d}t}(\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))) + \omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t)) + \dot{\gamma}(t)x(t)&\\ - \Big(\frac{1}{\omega} - \gamma(t)\Big)\dot{x}(t) + \frac{1}{\omega}\left(\Big(\frac{1}{\omega} - \gamma(t) - \dot{\gamma}(t)\omega\Big)x(t) - \frac{1}{\omega}y(t)\right) &= 0,\end{aligned}$$

and using again the first equation in (37) to eliminate $y(t)$, we get:

$$\begin{aligned}&\ddot{x}(t) + \omega\frac{\mathrm{d}}{\mathrm{d}t}(\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))) + \omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t)) + \dot{\gamma}(t)x(t) - \Big(\frac{1}{\omega} - \gamma(t)\Big)\dot{x}(t)\\ &+\frac{1}{\omega}\left(\Big(\frac{1}{\omega} - \gamma(t) - \dot{\gamma}(t)\omega\Big)x(t) + \dot{x}(t) + \omega\Big(\nabla G_{\mathsf{m}_{\bar{x}}}x(t) + \mathbf{e}_{\bar{x}}(x(t))\Big) - \Big(\frac{1}{\omega} - \gamma(t)\Big)x(t)\right) = 0,\end{aligned}$$

and after simplifications, we recover (ISEHD$_{\gamma,\omega}$). Conversely, let x be a trajectory solution to (ISEHD$_{\gamma,\omega}$) with initial conditions $(x_0, v_0) \in \mathcal{H}\times\mathcal{H}$ and define

$$y(t) = -\omega\left(\dot{x}(t) + \omega\Big(\nabla G_{\mathsf{m}_{\bar{x}}}x(t) + \mathbf{e}_{\bar{x}}(x(t))\Big) - \Big(\frac{1}{\omega} - \gamma(t)\Big)x(t)\right).$$

By differentiating the previous formula and using (ISEHD$_{\gamma,\omega}$), we recover the second equation of (37), as desired. ■

In the above, we have taken the derivatives in time as if they exist. We will show shortly that this is actually the case.

4.1.2. *Existence and uniqueness of a solution trajectory.* Proposition 2 opens the door to considering even the nonsmooth version of (ISEHD$_{\gamma,\omega}$) via (37) requiring only that $g \in \Gamma_0(\mathcal{H})$. Thus, writing $\partial G_{\mathsf{m}_{\bar{x}}}(x(t)) = \partial g + \mathbb{E}_{\xi\sim\mathsf{m}_{\bar{x}}}(\nabla f(\cdot,\xi))$, we can consider the differential inclusion

$$\begin{cases}\dot{x}(t) + \omega\Big(\partial G_{\mathsf{m}_{\bar{x}}}(x(t)) + \mathbf{e}_{\bar{x}}(x(t))\Big) - \left(\frac{1}{\omega} - \gamma(t)\right)x(t) + \frac{1}{\omega}y(t) & \ni 0\\ \dot{y}(t) - \left(\frac{1}{\omega} - \gamma(t) - \dot{\gamma}(t)\omega\right)x(t) + \frac{1}{\omega}y(t) & = 0,\end{cases} \tag{38}$$

with initial conditions $(x(t_0) = x_0, y(t_0) = y_0) \in \mathrm{dom}(G_{\mathsf{m}_{\bar{x}}}) \times \mathcal{H}$.

One of the main advantages of (38) is that it can be easily recast as a differential inclusion governed by a Lipschitz perturbation of a maximal monotone operator on the product space $\mathcal{H}\times\mathcal{H}$. Indeed, setting $Z(t) = (x(t), y(t))$, $\mathcal{A}(x,y) = (\omega\partial G_{\mathsf{m}_{\bar{x}}}(x), 0)$ and

$$\mathcal{E}(t,x,y) = \left(\omega\mathbf{e}_{\bar{x}}(x(t)) - \Big(\frac{1}{\omega} - \gamma(t)\Big)x(t) + \frac{1}{\omega}y(t), -\Big(\frac{1}{\omega} - \gamma(t) - \dot{\gamma}(t)\omega\Big)x(t) + \frac{1}{\omega}y(t)\right),$$

we immediately see that (38) can be written as

$$\dot{Z}(t) + \mathcal{A}(Z(t)) + \mathcal{E}(t, Z(t)) \ni 0_{\mathcal{H}\times\mathcal{H}}, \; Z(0) = (x_0, y_0). \tag{39}$$

Observe that $\mathcal{A}$ is nothing but the subdifferential of the function $H(x,y) = \omega G_{\mathsf{m}_{\bar{x}}}(x)$, and thus $\mathcal{A}$ is maximal monotone if the smooth part is monotone (see Lemma 3). Then, (38) can be written as the differential inclusion in the product space $\mathcal{H}\times\mathcal{H}$

$$\begin{cases}\dot{Z}(t) + \mathcal{A}(Z(t)) + \mathcal{E}(t, Z(t)) \ni 0_{\mathcal{H}\times\mathcal{H}}, \; \text{a.e } t \in [t_0, T]\\ Z(0) = (x_0, y_0) \in \mathrm{dom}(g)\times\mathcal{H},\end{cases} \tag{MIS}$$

which fits in the framework of Lipschitz perturbations of maximal monotone operators as in Section 3. Notice that this formulation is different from the classical Hamiltonian one.

Before stating the main result, let us recall that we endow the product space with the scalar product $\langle (u,v),(u^*,v^*)\rangle_{\mathcal{H}\times\mathcal{H}} = \langle u,u^*\rangle + \langle v,v^*\rangle$, and the induced norm $\|(u,v)\|_{\mathcal{H}\times\mathcal{H}} = \sqrt{\|u\|^2 + \|v\|^2}$. We have the following auxiliary results

Lemma 5. *Suppose that Assumptions 2 and 6 hold. Consider the operator $\mathcal{E}$ where $\omega > 0$ and $\gamma \in C^1([t_0,+\infty[)$. Then, for any $t \in [t_0,T]$, $\mathcal{E}(t,\cdot,\cdot)$ is Lipschitz continuous on $\mathcal{H}\times\mathcal{H}$.*

The dependence on t of the Lipschitz constant appears only through $|\gamma(t)|$ and $|\dot\gamma(t)|$. With the two most popular choices of $\gamma(t)$, constant (as in heavy ball method) or the asymptotically vanishing viscous damping $\gamma(t) = \alpha/t$, both $|\gamma(t)|$ and $|\dot\gamma(t)|$ are uniformly bounded. This makes $\mathcal{E}(t,\cdot,\cdot)$ Lipschitz continuous uniformly in $t > t_0$.

Proof. Let $u, v, u^*, v^* \in \mathcal{H}$ and set $p = (\frac{1}{\omega} - \gamma(t)), q = (\frac{1}{\omega} - \gamma(t) - \dot\gamma(t)\omega)$. We have, for $t \in [t_0,T]$

$$
\begin{aligned}
&\|\mathcal{E}(t,u,v) - \mathcal{E}(t,u^*,v^*)\|_{\mathcal{H}\times\mathcal{H}} \\
&= \left\| \Big(\omega(\mathbf{e}_{\bar x}(u) - \mathbf{e}_{\bar x}(u^*) + p(u^*-u) + \frac{1}{\omega}(v-v^*), \frac{1}{\omega}(v-v^*) + q(u^*-u) \Big) \right\|_{\mathcal{H}\times\mathcal{H}} \\
&= \sqrt{\|(\omega(\mathbf{e}_{\bar x}(u) - \mathbf{e}_{\bar x}(u^*) + p(u^*-u) + \frac{1}{\omega}(v-v^*)\|^2 + \|\frac{1}{\omega}(v-v^*) + q(u^*-u)\|^2} \\
&\leq \sqrt{2\omega^2\|\mathbf{e}_{\bar x}(u) - \mathbf{e}_{\bar x}(u^*)\|^2 + (4p^2+2q^2)\|u-u^*\|^2 + \frac{6}{\omega^2}\|v-v^*\|^2} \\
&= \sqrt{(2\omega^2(2L+\beta\tau)^2 + 4p^2 + 2q^2)\|u-u^*\|^2 + \frac{6}{\omega^2}\|v-v^*\|^2} \\
&\leq \left(\sqrt{2}\omega(2L+\beta\tau) + 2|p| + \sqrt{2}|q| + \frac{\sqrt{6}}{\omega}\right)\|(u,v) - (u^*,v^*)\|_{\mathcal{H}\times\mathcal{H}} \\
&= K(\omega,\beta,\tau,\gamma,t)\|(u,v) - (u^*,v^*)\|_{\mathcal{H}\times\mathcal{H}}
\end{aligned}
$$

where we have used Young's inequality and Lemma 4 for the Lipschitz continuity of the operator $\mathbf{e}_{\bar x}$. ∎

Proposition 3. *Assume that Assumptions 2, 6 and 7 hold where we only require $g \in \Gamma_0(\mathcal{H})$ (but possibly nonsmooth). Suppose also that $\omega > 0$ and $\gamma \in C^1([t_0,+\infty[;\mathcal{H})$ such that both γ and $\dot\gamma \in L^2([t_0,T])$ for all $T > t_0$. Then, for any initial data $x_0 \in \mathrm{dom}(g)$ and $y_0 \in \mathcal{H}$, there exists a unique global strong solution $(x,y) : [t_0,+\infty[\to \mathcal{H}\times\mathcal{H}$ to (MIS) such that $x(t_0) = x_0$ and $y(t_0) = y_0$. Moreover, the solution $Z = (x,y)$ satisfies the following properties*

- *(i) $y \in C^1([t_0,+\infty[;\mathcal{H})$, and $\dot y(t) - \left(\frac{1}{\omega} - \gamma(t) - \omega\dot\gamma(t)\right)x(t) + \frac{1}{\omega}y(t) = 0$, for all $t \geq t_0$;*
- *(ii) x is absolutely continuous on $[t_0,T]$ and $\dot x \in L^2(t_0,T;\mathcal{H})$ for all $T > t_0$;*
- *(iii) $x(t) \in \mathrm{dom}(\partial g)$ for all $t > t_0$;*
- *(iv) x is Lipschitz continuous on any compact subinterval of $]t_0,+\infty[$;*
- *(v) the function $t \in [t_0,+\infty[\mapsto G_{\mathsf{m}_{\bar x}}(x(t))$ is absolutely continuous on $[t_0,T]$ for all $T > t_0$;*
- *(vi) there exists a function $\xi : [t_0,+\infty[\to \mathcal{H}$ such that*
  - *(a) $\xi(t) \in \partial G_{\mathsf{m}_{\bar x}}(x(t))$ for all $t > t_0$;*
  - *(b) $\dot x(t) + \omega\xi(t) - \left(\frac{1}{\omega} - \gamma(t)\right)x(t) + \frac{1}{\omega}y(t) = 0$ for almost every $t > t_0$;*
  - *(c) $\xi \in L^2(t_0,T;\mathcal{H})$ for all $T > t_0$;*
  - *(d) $\frac{\mathrm{d}}{\mathrm{d}t}G_{\mathsf{m}_{\bar x}}(x(t)) = \langle \xi(t), \dot x(t)\rangle$ for almost every $t > t_0$.*

Proof. In view of our assumptions, $\mathcal{A}$ is maximal monotone (see Lemma 3) and $\mathcal{E}$ is Lipschitz continuous thanks to Lemma 5. We deduce the existence of a unique strong global solution $Z = (x, y) : [t_0, T] \to \mathcal{H} \times \mathcal{H}$ of (MIS) via [26, Proposition 3.12]. The verification of items (iii)-(vi) can be done by following the main arguments of [19, Theorem 4.4] ■

Assuming that $g \in C^1(\mathcal{H})$ with a Lipschitz continuous gradient, the conclusions of Proposition 3 can be strengthened as follows.

Proposition 4. *Assume that Assumptions 2, 6 and 7 hold with ∇g Lipschitz continuous. Suppose also that $\omega > 0$ and $\gamma \in C^1([t_0, +\infty[; \mathcal{H})$ such that both γ and $\dot{\gamma} \in L^2([t_0, T])$ for all $T > t_0$. Then, for any $t_0 > 0$, and any Cauchy data $(x_0, \dot{x}_0)$, the system* (ISEHD$_{\gamma,\omega}$) *admits a unique global solution $x \in C^1([t_0, +\infty[; \mathcal{H})$ satisfying $(x(t_0), \dot{x}(t_0)) = (x_0, \dot{x}_0)$. Moreover, $\dot{x}$, $\nabla G_{\mathsf{m}_{\bar{x}}}(x(\cdot))$ and $\mathbf{e}_{\bar{x}}(x(\cdot))$ are absolutely continuous.*

Proof. Under our regularity assumptions, we get from the conclusion of Proposition 3 and the first equation of (37) that x is a $C^1([t_0, +\infty[; \mathcal{H})$ function. Moreover, $\dot{x}$ is absolutely continuous thanks to Lipschitz continuity of $\nabla G_{\mathsf{m}_{\bar{x}}}$ and $\mathbf{e}_{\bar{x}}$, absolute continuity of x already established in Proposition 3, together with continuity of γ, $\dot{\gamma}$ and y. ■

Remark 4.1. The lack of differentiability of $\mathbf{e}_{\bar{x}}$ is the main obstacle for showing the existence of a classical solution even if $G_{\mathsf{m}_{\bar{x}}}$ is assumed C^2. Moreover, arguing using the non-autonomous Cauchy-Lipschitz theorem, the convexity assumption on g can be dropped in Proposition 4 at the price of assuming γ to be twice continuously differentiable on $[t_0, +\infty[$. The proof is left to the reader.

4.2. **Convergence properties.** Now let us examine the convergence properties of (ISEHD$_{\gamma,\omega}$). We will work under Assumptions 2, 6 and 7 where ∇g is Lipschitz continuous. The μ- strong convexity of $G_{\mathsf{m}_{\bar{x}}}$ allows to tune the viscous damping coefficient to the modulus μ by taking $\gamma(t) \equiv 2\sqrt{\mu}$ as advocated in [9, 14]. From now on, we focus on the following system

$$\ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)) + \omega \frac{\mathrm{d}}{\mathrm{d}t}\left(\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\right) + \mathbf{e}_{\bar{x}}(x(t)) + \omega \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t)) = 0. \tag{ISEHD$_{2\sqrt{\mu},\omega}$}$$

Thanks to our assumptions, this systems falls within the framework of the previous section so that Proposition 4 applies.

To perform Lyapunov analysis, let us define the following energy function $\mathcal{V} : [t_0, \infty[\to \mathbb{R}^+$ by

$$\mathcal{V}(t) := G_{\mathsf{m}_{\bar{x}}}(x(t)) - G^*_{\mathsf{m}_{\bar{x}}} + \frac{1}{2}\|v(t)\|^2, \text{ where } v(t) := \sqrt{\mu}\,(x(t) - \bar{x}) + \dot{x}(t) + \omega \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)).$$

where $G^*_{\mathsf{m}_{\bar{x}}} = \min G_{\mathsf{m}_{\bar{x}}}(\mathcal{H})$.

We prove the following result.

Theorem 3. *Assume that Assumptions 2, 6 and 7 hold and ∇g is Lipschitz continuous. Let $x : [t_0, +\infty[\to \mathcal{H}$ be the solution trajectory of* (ISEHD$_{2\sqrt{\mu},\omega}$). *Suppose that $\rho := \frac{\beta\tau}{\mu}$ and the damping coefficient ω satisfy*

$$0 \le \omega \le \min\left(\frac{1}{2\sqrt{\mu}}, \frac{\sqrt{\mu}}{2\sqrt{2}(2L + \beta\tau)}\right) \quad \textit{and} \quad 16\rho^2 + \omega < 1. \tag{40}$$

Then, we have:

(1) for all $t \geq t_0$

$$\mathcal{V}(t) \leq \mathcal{V}(t_0)e^{-\frac{\sqrt{\mu}}{4}(t-t_0)}.$$

In particular

$$\frac{\mu}{2}\|x(t)-\bar{x}\|^2 \leq G_{\mathsf{m}_{\bar{x}}}(x(t)) - G^*_{\mathsf{m}_{\bar{x}}} \leq \mathcal{V}(t_0)e^{-\frac{\sqrt{\mu}}{4}(t-t_0)}.$$

(2) There exists $C > 0$ *such that,*

$$e^{-\sqrt{\mu}t}\int_{t_0}^{t} e^{\sqrt{\mu}s}\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(s))\|^2 \mathrm{d}s \leq Ce^{-\frac{\sqrt{\mu}}{4}t}, \ \forall t \geq t_0.$$

Some remarks are in order before proving this result.

Remark 4.2.

- Compared to the convergence results in Theorem 2 on the first-order system (SMI), Theorem 3 not only provides a fast convergence rate of the trajectory and objective value, but also of the gradient. Observe also the $\sqrt{\mu}$ in the rate in Theorem 3 instead of μ Theorem 2. This recovers the known result that the system (ISEHD$_{2\sqrt{\mu},\omega}$) is faster than (SMI) for badly conditioned objectives, approaching the optimal rate of the class of strongly convex functions with Lipschitz continuous gradient.
- Contrary to [14], no assumption on the integrability of $\mathbf{e}_{\bar{x}}(x(\cdot))$ and $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(\cdot))$ is needed in Theorem 3. In fact, thanks to Corollary 1 and Lemma 4, the norms of these error terms will be absorbed in the right hand side of (44).
- In case $\omega = 0$, i.e., when the inertial dynamic is considered only with the viscous damping coefficient (which can be view as a perturbed HBF method), the condition (40) reduces to $\rho < \frac{1}{4}$. This contrasts with the convergence condition for first-order dynamics (cf. Theorem 2), where convergence is ensured in parameter regime $\rho < 1$. One possible explanation for this difference, is the potential occurrence of oscillations typical of inertial system, which may necessitate a stricter compatibility condition between the parameters τ, β, μ.

Proof. In view of our assumptions, Proposition 4 applies and we get that $v : [t_0, +\infty[\to \mathcal{H}$ is absolutely continuous and thus so is $\mathcal{V}$. We then have

$$\begin{aligned}\dot{\mathcal{V}}(t) &= \langle \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), \dot{x}(t)\rangle + \langle v(t), \sqrt{\mu}\dot{x}(t) + \ddot{x}(t) + \omega\frac{\mathrm{d}}{\mathrm{d}t}\left(\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\right)\rangle \\ &= \langle \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), \dot{x}(t)\rangle + \langle v(t), -\sqrt{\mu}\dot{x}(t) - \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)) - \mathbf{e}_{\bar{x}}(x(t)) - \omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t))\rangle,\end{aligned} \tag{41}$$

where we used the constitutive equation in (ISEHD$_{2\sqrt{\mu},\omega}$). Replacing v by its expression and rearranging, we arrive at

$$\begin{aligned}&\dot{\mathcal{V}}(t) + \mu\langle \dot{x}(t), x(t)-\bar{x}\rangle + \sqrt{\mu}\|\dot{x}(t)\|^2 + \sqrt{\mu}\langle \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), x(t)-\bar{x}\rangle \\ &\quad + \omega\sqrt{\mu}\langle \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), \dot{x}(t)\rangle + \omega\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|^2 = -\langle v(t), \mathbf{e}_{\bar{x}}(x(t)) + \omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t))\rangle.\end{aligned} \tag{42}$$

Using μ-strong convexity of $G_{\mathsf{m}_{\bar{x}}}$, we have

$$\langle \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), x(t)-\bar{x}\rangle \geq G_{\mathsf{m}_{\bar{x}}}(x(t)) - G^*_{\mathsf{m}_{\bar{x}}} + \frac{\mu}{2}\|x(t)-\bar{x}\|^2, \tag{43}$$

and plugging this into (42), we get

$$\dot{\mathcal{V}}(t) + \sqrt{\mu}\Theta(t) \leq \|v(t)\|\|\mathbf{e}_{\bar{x}}(x(t)) + \omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t))\|, \tag{44}$$

where

$$\Theta(t) := G_{\mathsf{m}_{\bar{x}}}(x(t)) - G^*_{\mathsf{m}_{\bar{x}}} + \frac{\mu}{2}\|x(t)-\bar{x}\|^2 + \sqrt{\mu}\langle \dot{x}(t), x(t)-\bar{x}\rangle + \|\dot{x}(t)\|^2 \\ + \omega\langle \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), \dot{x}(t)\rangle + \frac{\omega}{\sqrt{\mu}}\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|^2.$$

Using the definition of $\mathcal{V}(t)$, we may rewrite $\Theta(t)$ as

$$\Theta(t) = \mathcal{V}(t) + \frac{1}{2}\|\dot{x}(t)\|^2 - \omega\sqrt{\mu}\langle \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), x(t)-\bar{x}\rangle + \left(\frac{\omega}{\sqrt{\mu}} - \frac{\omega^2}{2}\right)\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|^2.$$

Consequently, (44) becomes

$$\dot{\mathcal{V}}(t) + \sqrt{\mu}\mathcal{V}(t) + \frac{\sqrt{\mu}}{2}\|\dot{x}(t)\|^2 \\ + \sqrt{\mu}\left(\left(\frac{\omega}{\sqrt{\mu}} - \frac{\omega^2}{2}\right)\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|^2 - \omega\sqrt{\mu}\langle \nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), x(t)-\bar{x}\rangle\right) \\ \leq \|v(t)\|\|\mathbf{e}_{\bar{x}}(x(t)) + \omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t))\|. \quad (45)$$

Using strong convexity of $G_{\mathsf{m}_{\bar{x}}}$ again, and discarding the quadratic term in $v(t)$, we have

$$\mathcal{V}(t) = \frac{1}{2}\mathcal{V}(t) + \frac{1}{2}\mathcal{V}(t) \geq \frac{1}{2}\mathcal{V}(t) + \frac{\mu}{4}\|x(t)-\bar{x}\|^2.$$

Observing that $\frac{\omega}{2\sqrt{\mu}} \leq \frac{\omega}{\sqrt{\mu}} - \frac{\omega^2}{2}$ for $0 \leq \omega \leq \frac{1}{\sqrt{\mu}}$, we end up with

$$\dot{\mathcal{V}}(t) + \frac{\sqrt{\mu}}{2}\mathcal{V}(t) + \frac{\sqrt{\mu}}{2}\|\dot{x}(t)\|^2 \\ + \sqrt{\mu}\left(\frac{\mu}{4}\|x(t)-\bar{x}\|^2 + \frac{\omega}{2\sqrt{\mu}}\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|^2 - \omega\sqrt{\mu}\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|\|x(t)-\bar{x}\|\right) \\ \leq \|v(t)\|\|\mathbf{e}_{\bar{x}}(x(t)) + \omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t))\|. \quad (46)$$

Now let us treat the right hand side of this inequality. Thanks to Corollary 1, we have $\|\mathbf{e}_{\bar{x}}(x(t))\| = \|\mathbf{e}_{\bar{x}}(x(t)) - \mathbf{e}_{\bar{x}}(\bar{x})\| \leq \beta\tau\|x(t)-\bar{x}\|$, and by Lemma 4 $\|\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t))\| \leq (2L+\beta\tau)\|\dot{x}(t)\|$. Since $\mathcal{V}(t) \geq \frac{1}{2}\|v(t)\|^2$, applying Young's inequality yields

$$\|v(t)\|\|\mathbf{e}_{\bar{x}}(x(t)) + \omega\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t))\| \leq \frac{\sqrt{\mu}}{8}\|v(t)\|^2 + \frac{4}{\sqrt{\mu}}\|\mathbf{e}_{\bar{x}}(x(t))\|^2 + \frac{4\omega^2}{\sqrt{\mu}}\|\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_{\bar{x}}(x(t))\|^2 \\ \leq \frac{\sqrt{\mu}}{4}\mathcal{V}(t) + \frac{4\beta^2\tau^2}{\sqrt{\mu}}\|x(t)-\bar{x}\|^2 + \frac{4\omega^2(2L+\beta\tau)^2}{\sqrt{\mu}}\|\dot{x}(t)\|^2 \quad (47)$$

We get after rearranging the terms

$$\dot{\mathcal{V}}(t) + \frac{\sqrt{\mu}}{4}\mathcal{V}(t) + \left(\frac{\sqrt{\mu}}{2} - \frac{4\omega^2(2L+\beta\tau)^2}{\sqrt{\mu}}\right)\|\dot{x}(t)\|^2 + \sqrt{\mu}\ \Psi(t) \leq 0, \quad (48)$$

where

$$\Psi(t) = \left(\frac{\mu}{4} - \frac{4\beta^2\tau^2}{\mu}\right)\|x(t)-\bar{x}\|^2 + \frac{\omega}{2\sqrt{\mu}}\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|^2 - \omega\sqrt{\mu}\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|\|x(t)-\bar{x}\|.$$

Setting

$$\mathbf{a} = \frac{\mu}{4} - \frac{4\beta^2\tau^2}{\mu}, \mathbf{b} = \frac{\omega}{2\sqrt{\mu}}, \mathbf{c} = -\frac{\omega\sqrt{\mu}}{2}$$

and $X = \|x(t) - \bar{x}\|$ and $Y = \|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|$, we see that Ψ can be written as a quadratic form $\mathcal{Q} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ with $\mathcal{Q}(X,Y) = \mathbf{a}\|X\|^2 + 2\mathbf{c}\langle X, Y\rangle + \mathbf{b}\|Y\|^2$. By assumption $\mathbf{a},\mathbf{b} \geq 0$, and the discriminant $\mathbf{c}^2 - \mathbf{ab}$ of $\mathcal{Q}$ is nonpositive. Indeed, since $0 \leq \omega \leq \frac{1}{2\sqrt{\mu}}$

$$\mathbf{c}^2 - \mathbf{ab} = \frac{\omega^2\mu}{4} - \frac{\omega}{2\sqrt{\mu}}\left(\frac{\mu}{4} - \frac{4\beta^2\tau^2}{\mu}\right) \leq \frac{\omega^2\sqrt{\mu}}{8} - \frac{\omega\sqrt{\mu}}{2}\left(\frac{1}{4} - 4\rho^2\right) = \frac{\omega\sqrt{\mu}}{8}(\omega - 1 + 16\rho^2) < 0.$$

Hence, $\Psi(t) \geq 0$ and since $\omega \leq \frac{\sqrt{\mu}}{2\sqrt{2}(2L+\beta\tau)}$, we get $\left(\frac{\sqrt{\mu}}{2} - \frac{4\omega^2(2L+\beta\tau)^2}{\sqrt{\mu}}\right) \geq 0$, and thus

$$\dot{\mathcal{V}}(t) + \frac{\sqrt{\mu}}{4}\mathcal{V}(t) \leq 0,$$

which gives after integration

$$\mathcal{V}(t) \leq \mathcal{V}(t_0)e^{\frac{-\sqrt{\mu}}{4}(t-t_0)}. \tag{49}$$

Therefore, $\lim_{t\to\infty} \mathcal{V}(t) = 0$ and in particular

$$\lim_{t\to\infty} G_{\mathsf{m}_{\bar{x}}}(x(t)) - G^*_{\mathsf{m}_{\bar{x}}} = 0 \text{ and } \lim_{t\to\infty} \|v(t)\| = 0. \tag{50}$$

This implies, using strong convexity of $G_{\mathsf{m}_{\bar{x}}}$

$$\lim_{t\to\infty} \|x(t) - \bar{x}\| = 0,$$

which gives that

$$\lim_{t\to\infty} \|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\| = 0.$$

We deduce from (50) that $\lim_{t\to\infty} \|\dot{x}(t)\| = 0$.

Coming back to (49), we have, by definition of $\mathcal{V}$ and μ-strong convexity of $G_{\mathsf{m}_{\bar{x}}}$, that

$$\frac{\mu}{2}\|x(t) - \bar{x}\|^2 \leq G_{\mathsf{m}_{\bar{x}}}(x(t)) - G^*_{\mathsf{m}_{\bar{x}}} \leq \mathcal{V}(t_0)e^{-\frac{\sqrt{\mu}}{4}(t-t_0)} \text{ and } \|v(t)\|^2 \leq 2\mathcal{V}(t_0)e^{-\frac{\sqrt{\mu}}{4}(t-t_0)}. \tag{51}$$

Developing in (51) we have

$$\begin{aligned}\mu\|x(t) - \bar{x}\|^2 + \|\dot{x}(t)\|^2 + \omega^2\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|^2 + 2\omega\sqrt{\mu}\langle\nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), x(t) - \bar{x}\rangle \\ + 2\omega\langle\nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), \dot{x}(t)\rangle + 2\sqrt{\mu}\langle\dot{x}(t), x(t) - \bar{x}\rangle \leq Ce^{-\frac{\sqrt{\mu}}{4}t},\end{aligned} \tag{52}$$

where $C = 2\mathcal{V}(t_0)e^{\frac{\sqrt{\mu}}{4}t_0}$.

Since $\langle\nabla G_{\mathsf{m}_{\bar{x}}}(x(t)), x(t) - \bar{x}\rangle \geq G_{\mathsf{m}_{\bar{x}}}(x(t)) - G^*_{\mathsf{m}_{\bar{x}}}$, we deduce from (52) that

$$\dot{U}(t) + \sqrt{\mu}U(t) + \omega^2\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(t))\|^2 \leq Ce^{-\frac{\sqrt{\mu}}{4}t}, \tag{53}$$

where $U(t) := \sqrt{\mu}\|x(t) - \bar{x}\|^2 + 2\omega\left(G_{\mathsf{m}_{\bar{x}}}(x(t)) - G^*_{\mathsf{m}_{\bar{x}}}\right)$. Integrating (53), we obtain after elementary computation

$$e^{-\sqrt{\mu}t}\int_{t_0}^{t} e^{\sqrt{\mu}s}\|\nabla G_{\mathsf{m}_{\bar{x}}}(x(s))\|^2 ds \leq C_1 e^{-\frac{\sqrt{\mu}}{4}t},$$

as desired. ■

We end this section with a similar result to Corollary 2, which is a direct consequence of Assumption 2 and Theorem 3(1).

Corollary 3. *Let $x : [t_0, +\infty[\to \mathcal{H}$ be the solution of* (ISEHD$_{2\sqrt{\mu},\omega}$) *where (40) holds. Then* $\forall t \geq t_0$

$$\mathrm{W}_1(\mathsf{m}_{x(t)}, \mathsf{m}_{\bar{x}}) \leq \tau\sqrt{\frac{2}{\mu}\mathcal{V}(t_0)e^{-\frac{\sqrt{\mu}}{8}(t-t_0)}}.$$

5. On coarse Ricci curvature

In this section we discuss some dynamical and geometrical properties of the family $(\mathsf{m}_x)_{x\in\mathcal{H}}$, particularly the notion of Ollivier-Ricci curvature and how it it tightly related to Assumption 2. In fact, the family of probabilities $\mathsf{m} = (\mathsf{m}_x)_{x\in\mathcal{H}}$ and its Lipschitz behavior with respect to the $\mathbb{W}_1$-Wasserstein distance, reveals that a natural setting to address monotone inclusions of the form (9), and thus stochastic optimization problems with decision-dependent distributions is the framework of *metric random walk spaces* (see, e.g., [41, 44]). All definitions of this section can be found in [36, 41].

5.1. **Metric random walk spaces.** Before going further, let us recall the following definitions to introduce a couple of probabilistic notions.

Definition 6 (Random walks [44])**.** Given a Polish space (X, d). A family of probabilities $\mathsf{m} = (\mathsf{m}_x)_{x\in X}$ is a random walk on X if $\mathsf{m}_x \in \mathcal{P}(\Xi)$ for each $x \in X$ and

- m_x depends measurably on $x \in X$,
- Each m_x has finite first-order moment, i.e., for some $x^o \in X$, $\mathbb{E}_{y\sim\mathsf{m}_x}\mathrm{d}(y, x^o) < \infty$.

Then (X, d) equipped with a random walk m is a metric random walk space (m.r.w.s for short), and we denote it by $[X, \mathrm{d}, \mathsf{m}]$.

Let us recall the notion of invariant and ergodic measures.

Definition 7 (Invariance)**.** Let ν be a σ-finite measure on X and m a random walk on $(X, \mathcal{B})$. We say that ν is invariant with respect to m if $\nu \star \mathsf{m} = \nu$, where $\nu \star \mathsf{m}$ is the convolution of ν by the random walk m and is defined by

$$\nu \star \mathsf{m}(A) = \int_X \mathsf{m}_x(A)\mathrm{d}\nu(x) \text{ for all } A \in \mathcal{B}.$$

As pointed out in [44], each measure m_x can be seen as a replacement of a sphere around x. While in a probabilistic framework one think about a Markov chain whose transition kernel from x to y in n steps is defined by

$$\mathrm{d}\mathsf{m}_x^{*n}(y) = \int_{z\in X} \mathrm{d}\mathsf{m}_x^{*(n-1)}(z)\mathrm{d}\mathsf{m}_z(y), \tag{54}$$

with $\mathsf{m}_x^1 = \mathsf{m}_x$ and $\mathsf{m}_x^0 = \delta_x$.

In the sequel, we assume that $(\mathcal{H}, \mathrm{d})$ is a separable real Hilbert space, and thus a Polish space, where $\mathrm{d}(x, y) = \langle x - y, x - y\rangle^{1/2}$.

5.2. **Feller Property.** Let us recall the following definition.

Definition 8. We say that $\mathsf{m} := (\mathsf{m}_x)_x$ has the weak-Feller property if and only if for every sequence $x_n \to x^0 \in \mathcal{H}$ we have $\mathsf{m}_{x_n} \rightharpoonup \mathsf{m}_{x^o}$, i.e., $\int f\mathrm{d}\mathsf{m}_{x_n} \to \int f\mathrm{d}\mathsf{m}_{x^o}$ for any $f \in C_b(\mathcal{H})$.

It turns that Assumption 2 implies directly that the family m is weak-Feller.

Proposition 5. *Under Assumption 2,* m *has the weak-Feller property. Moreover, for each* $x \in \mathcal{H}$*,* m_x *has finite first-order moment.*

Proof. Let $x^o \in \mathcal{H}$ and $(x_n)_n$ a sequence of $\mathcal{H}$ such that $x_n \to x^o$ as $n \to 0$. Then Assumption 2 gives

$$\mathbb{W}_1(\mathsf{m}_{x_n}, \mathsf{m}_{x^o}) \le \tau\|x_n - x^o\|,$$

and thus $\lim_{n\to\infty} \mathbb{W}_1(\mathsf{m}_{x_n}, \mathsf{m}_{x^o}) = 0$. Thanks to [4, Proposition 7.1.5], (m_{x_n}) has uniformly integrable p-moments with $p \ge 1$ and narrowly converges towards m_{x^o}. In particular m is weak-Feller and each m_x has finite first-order moments. ∎

Remark 5.1. We already know that $\mathsf{m}_x \in \mathcal{P}(\Xi)$ for each $x \in \mathcal{H}$ and that $x \mapsto \mathsf{m}_x(C)$ is measurable for each $C \in \mathcal{B}$. Moreover, thanks to Proposition 5, we have finiteness of first-order moments of each m_x, so that the family m satisfies the requirements of Definition 6. This shows that a natural setting to address dynamics of the form (9) is the metric random walk space $[\mathcal{H}, \mathrm{d}, \mathsf{m}]$. Many diffusion and variational problems has been studies within this framework, with allows in particular consider nonlocal continuum problems or problems on weighted graphs (see, e.g., [41] and the references therein).

Remark 5.2. Let us point out that if υ is an invariant measure with respect to m then it is also and invariant measure with respect to m^{*n} for every $n \in \mathbb{N}$, where m^{*n} is the n-step transition probability function given by (54). It turns out that weak-Feller property implies that every weak$-*$ limit υ of $(\mathsf{m}^{*n})_n$ is an invariant measure of m cf. [36, Proposition 7.2.2] (see also [33, Proposition 12.3.4]). However, without assuming at first the existence of an invariant measure with respect to m, the measure υ may be trivial. Without further compactness assumptions on the metric space (see, e.g., [36, Theorem 7.2.3]) one needs some Lyapunov like condition to ensure the existence of an invariant measure υ of the weak-Feller family m (see, e.g., [36, Theorem 7.2.4] or [33, Theorem 12.3.3]). As we will see in Corollary 5, another way to obtain the existence of invariant measures is having a positive lower bound on the coarse Ricci curvature of $[\mathcal{H}, \mathrm{d}, \mathsf{m}]$.

5.3. **Ollivier-Ricci curvature.** Let us discuss here the connexion between Assumption 2 and the so-called coarse or Olliver-Ricci curvature (ORC for short). The results can be found in [44] or [45]. A more recent presentation can be found in [41].

Definition 9 (Ollivier-Ricci curvature [44])**.** Let $[\mathcal{H}, \mathrm{d}, \mathsf{m}]$ be a m.r.w.s. Then, for any distinct points $x, y \in \mathcal{H}$, the ORC along (x, y) is defined as:

$$\kappa_{\mathsf{m}}(x,y) = 1 - \frac{\mathbb{W}_1(\mathsf{m}_x, \mathsf{m}_y)}{\mathrm{d}(x,y)}, \tag{55}$$

The ORC of $[\mathcal{H}, \mathrm{d}, \mathsf{m}]$ is defined as

$$\kappa_{\mathsf{m}} := \inf_{x \neq y} \kappa_{\mathsf{m}}(x,y). \tag{56}$$

We clearly see from (55) that $\kappa_{\mathsf{m}}(x,y) \leq 1$. Moreover, rearranging the terms, we have

$$\mathbb{W}_1(\mathsf{m}_x, \mathsf{m}_y) = (1 - \kappa_{\mathsf{m}}(x,y))\,\mathrm{d}(x,y). \tag{57}$$

Consequently, having some lower bound $\kappa_{\mathsf{m}}(x,y) \geq c \in \mathbb{R}$ for any $x, y \in \mathcal{H}$ gives

$$\mathbb{W}_1(\mathsf{m}_x, \mathsf{m}_y) \leq (1-c)\mathrm{d}(x,y), \tag{58}$$

which describes a Lipschitz behavior of the random walk m. This has to be compared to Assumption 2. Indeed, we see from Assumption 2 that, for $x \neq y$

$$1 - \tau \leq \kappa_{\mathsf{m}}(x,y) \leq 1, \tag{59}$$

so according to the values of τ we have different regimes on the ORC κ_{m} (cf. Table 1).

Table 1: Relation between the values of τ and κ_{m}.

Values of τ	0	< 1	≤ 1
Values of κ_{m}	1	$]0,1]$	$[0,1]$

Notice that Assumption 2 excludes both the cases $\kappa_{\mathsf{m}} \equiv 1$ and $\kappa_{\mathsf{m}} < 0$. Typically, $\tau = 0$, would give that $\kappa_{\mathsf{m}} = 1$ in other words $\mathbb{W}_1(\mathsf{m}_x, \mathsf{m}_y) = 0$ for any $x, y \in \mathcal{H}$, i.e., the distribution m is constant.

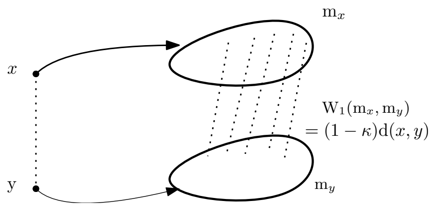
m_x
x
$\mathrm{W}_1(\mathrm{m}_x, \mathrm{m}_y)$
$= (1-\kappa)\mathrm{d}(x, y)$
y
m_y

Figure 1: Illustration of the ORC.

Moreover, it turns that there is equivalence between the lower bound on κ_{m} in (59) and the Lipschitz behavior (58). This is directly related to a $\mathbb{W}_1$-contraction property cf. [44, Proposition 20].

Proposition 6. *Let* m *be a random walk on* $(\mathcal{H}, \mathrm{d})$ *and assume that* m_x *has finite moment for all* $x \in \mathcal{H}$*. Then*

$$\kappa_{\mathsf{m}}(x,y) \geq c \in \mathbb{R},\ \forall x \neq y \iff \mathbb{W}_1(\nu_1 \star \mathsf{m}, \nu_2 \star \mathsf{m}) \leq (1-c)\mathbb{W}_1(\nu_1, \nu_2)\ \forall \nu_1, \nu_2 \in \mathcal{P}_1(\Xi).$$

In view of Proposition 6, taking $\nu_1 = \delta_x$ and $\nu_2 = \delta_y$ for $x \neq y$, and $c = 1 - \tau$, we get

$$\mathbb{W}_1(\mathsf{m}_x, \mathsf{m}_y) = \mathbb{W}_1(\delta_x \star \mathsf{m}, \delta_y \star \mathsf{m}) \leq \tau \mathbb{W}(\delta_x, \delta_y) = \tau \mathrm{d}(x, y)$$

which is exactly Assumption 2 since the above inequality is trivial for $x = y$.

In the case of positive curvature, this contraction result implies the existence of a unique invariant measure for the random walk m when the ORC is positive.

Corollary 4 ([41])**.** *Assume that* $\kappa_{\mathsf{m}}(x,y) \geq c > 0$ *for all* $x \neq y$*. Then, the random walk* m *has a unique invariant measure* $\upsilon \in \mathcal{P}_1(\Xi)$*. Moreover, for any* $\nu \in \mathcal{P}_1(\Xi)$

(1) $\mathbb{W}_1(\nu \star \mathsf{m}^{*n}, \upsilon) \leq (1-c)^n \mathbb{W}_1(\nu, \upsilon)$, $\forall n \in \mathbb{N}$.

(2) $\mathbb{W}_1(\mathsf{m}^{*n}, \upsilon) \leq \frac{(1-c)^n}{c} \mathbb{W}_1(\delta_x, \upsilon)$, $\forall n \in \mathbb{N}, \forall x \in \mathcal{H}$.

In our setting, the parameter regime $\tau < 1$ would define a positive ORC. Consequently, we have the following

Corollary 5. *Assume that* m *satisfies Assumption 2 with* $\tau < 1$*. Then, there exists a unique invariant measure* $\upsilon \in \mathcal{P}_1(\Xi)$ *with respect to* m*.*

6. Application: Inertial primal-dual algorithm

This section is devoted to the application of the developed results in Section 3 and Section 4 to the following class of saddle-point problems

$$\inf_{x \in \mathcal{H}} \sup_{y \in \mathcal{K}} \mathsf{f}_{\mathsf{m}_x}(x) + \mathsf{g}(x) + \langle y, \mathrm{K}x \rangle - \mathsf{h}(y) - \mathsf{r}_{\mathsf{m}_y}(y), \tag{60}$$

where $\mathrm{K} : \mathcal{H} \to \mathcal{K}$ is a bounded linear operator, $\mathcal{K}$ is a real Hilbert space. We equip $\mathcal{H} \times \mathcal{K}$ with the product space inner product and associated norm. In what follows, we make the following assumptions:

Assumption 8.

(a) $\mathsf{f}_{\mathsf{m}_x}(x) := \mathbb{E}_{\xi \sim \mathsf{m}_x}(f(x, \xi))$, $f \in C^1(\mathcal{H} \times \Xi)$, $f(x, \cdot)$ is m_x-measurable and $C^{1,1}_{\beta_{\mathsf{f}}}(\Xi)$ for every $x \in \mathcal{H}$, and $\mathsf{f}_{\mathsf{m}_x}(\cdot, \xi) \in C^{1,1}_{L_{\mathsf{f}}}(\mathcal{H})$ for every $\xi \in \Xi$;

(b) $\mathsf{r}_{\mathsf{m}_y}(y) := \mathbb{E}_{\zeta \sim \mathsf{m}_y}(\mathsf{r}(y, \zeta))$, with $\mathsf{r} \in C^1(\mathcal{K} \times \mathcal{Z})$, $\mathsf{r}(y, \cdot)$ is m_y-measurable and $C^{1,1}_{\beta_{\mathsf{r}}}(\mathcal{Z})$ for every $y \in \mathcal{K}$, and $\mathsf{r}_{\mathsf{m}_y}(\cdot, \zeta) \in C^{1,1}_{L_{\mathsf{r}}}(\mathcal{K})$ for every $\zeta \in \mathcal{Z}$;

(c) $\mathsf{h} \in \Gamma_0(\mathcal{H})$ and $\mathsf{g} \in \Gamma_0(\mathcal{K})$;

(d) $\mathsf{f}_{\mathsf{m}_x} + \mathsf{g}$ is μ_p-strongly convex for all $x \in \mathcal{H}$, and $\mathsf{r}_{\mathsf{m}_y} + \mathsf{h}$ is μ_d-strongly convex for all $y \in \mathcal{K}$.

(e) There exists $\tau > 0$ such that

$$\mathbb{W}_1(\mathsf{m}_x \otimes \mathsf{m}_y, \mathsf{m}_{x'} \otimes \mathsf{m}_{y'}) \leq \tau \|(x,y) - (x',y')\|_{\mathcal{H}\times\mathcal{K}}, \text{ for all } x, x' \in \mathcal{H}, y, y' \in \mathcal{K}.$$

$\beta_{\mathsf{f}}, \beta_{\mathsf{r}}, L_{\mathsf{f}}, L_{\mathsf{r}}, \mu_p, \mu_d$ are all positive constants.

Assumption 8(a)–(c) are to be compared to Assumptions 1 and 5. This will be made clearer when we will cast the saddle point problem (60) and an inclusion problem reminiscent of (9). In the same way, Assumption 8(d) will be sufficient for Assumption 3 to hold. (e) is the version of Assumption 2 on the product space $\mathcal{H} \times \mathcal{K}$.

Problems of the form (60) arise in many fields such as image and signal processing, machine learning and partial differential equations. In the deterministic case, i.e., f does not depend on the distribution m and $\mathsf{g} \equiv 0$, such problems were studied in [29]. Later on, extensions were addressed in several works (see, e.g., [30, 48, 51]). In [24], the authors studied a fully stochastic variant of (60), i.e., $\mathsf{f}(x) = \mathbb{E}_{\xi\sim\mathsf{m}}(f(x,\xi)), \mathsf{g}(x) = \mathbb{E}_{\xi\sim\mathsf{m}}(g(x,\xi))$ and $\mathrm{K} = \mathbb{E}_{\xi\sim\mathsf{m}}(K(\xi))$ for some suitable functions f, g and operators K. Yet, the distribution m does not depend on the state x. Problems of the form (60) in their full generality are difficult to tacle directly because if the the presence of (primal and dual) state-dependent distributions. To the best of our knowledge, general saddle point problems of the form (60) have not been addressed in the literature.

6.1. **Formulation as a monotone inclusion.** As seen in Section 3, the appropriate notion of solutions of (60) is that of equilibria. Thus, our aim is to find an equilibrium point $(\bar{x}, \bar{y})$, i.e., that $(\bar{x}, \bar{y})$ a solution of the static saddle point problem

$$\inf_{x\in\mathcal{H}} \sup_{y\in\mathcal{K}} \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x,y), \tag{61}$$

where we have defined the Lagrangian for the (primal and dual) probability measures $m^p \in \mathcal{P}(\Xi)$ and $m^d \in \mathcal{P}(\mathcal{Z})$

$$\mathcal{L}_{\mathsf{m}^p,\mathsf{m}^d}(x,y) := \mathsf{f}_{\mathsf{m}^p}(x) + \mathsf{g}(x) + \langle y, \mathrm{K}x\rangle - \mathsf{h}(y) - \mathsf{r}_{\mathsf{m}^d}(y).$$

If the set of saddle points is nonempty (typically under strong duality, see e.g., [23, Chapter 15]), then $(x^\star, y^\star)$ is a saddle point for (61) if and only if

$$\mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x^\star, y) \leq \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x^\star, y^\star) \leq \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x, y^\star) \quad \forall (x,y) \in \mathcal{H}\times\mathcal{K},$$

or equivalently, if the following optimality condition holds [23, Corollary 19.19][1]

$$\begin{cases} 0 \in \nabla \mathsf{f}_{\mathsf{m}_{\bar{x}}}(x^\star) + \partial \mathsf{g}(x^\star) + \mathrm{K}^* y^\star \\ 0 \in \nabla \mathsf{r}_{\mathsf{m}_{\bar{y}}}(y^\star) + \partial \mathsf{h}(y^\star) - \mathrm{K}x^\star. \end{cases} \tag{62}$$

where K^* is the adjoint operator of K. Therefore, $(\bar{x}, \bar{y})$ will be said to be an equilibrium of (60) if it is a solution of (62). Equivalently, this can be cast as the monotone inclusion problem

$$0_{\mathcal{H}\times\mathcal{K}} \in \mathbf{T}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x^\star, y^\star), \tag{63}$$

where

$$\mathbf{T}_{\mathsf{m}^p,\mathsf{m}^d} = \mathbf{A} + \mathbf{L} + \mathbf{B}_{\mathsf{m}^p,\mathsf{m}^d} : \mathcal{H}\times\mathcal{K} \rightrightarrows \mathcal{H}\times\mathcal{K},$$
$$\mathbf{A} = \begin{pmatrix} \partial\mathsf{g} & 0 \\ 0 & \partial\mathsf{h} \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} 0 & \mathrm{K}^* \\ -\mathrm{K} & 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{B}_{\mathsf{m}_x,\mathsf{m}_y} = \begin{pmatrix} \nabla \mathsf{f}_{\mathsf{m}^p} & 0 \\ 0 & \nabla \mathsf{r}_{\mathsf{m}^d} \end{pmatrix}. \tag{64}$$

[1] Here, only convexity is needed rather than Assumption 8(d).

Remark 6.1. The authors in [52] studied problems of the form

$$\min_{x\in X}\max_{y\in Y}\mathbb{E}_{\xi\sim\mathsf{m}_{(\bar{x},\bar{y})}}\phi(x,y,\xi),$$

in finite dimension where X, Y are compact sets and ϕ is a convex-concave function that plays the role of the Lagrangian in our case. Such problems fall into the scope of (61) where X and Y can be absorbed in the functions g and h through their respective indicator functions ι_X and ι_Y.

The form (64) brings back our problem to the general setting studied in Section 3. The following result shows that the operators $\mathbf{A}$ and $\mathbf{B}_{\mathsf{m}^p,\mathsf{m}^d}$ indeed verify all the appropriate properties required there.

Lemma 6. *Under Assumption 8, the following properties hold for any probability measures* $m^p \in \mathcal{P}(\Xi)$ *and* $m^d \in \mathcal{P}(\mathcal{Z})$*:*

(i) *The operator* $\mathbf{A}$ *is maximally monotone.*

(ii) *The operator* $\mathbf{B}_{\mathsf{m}^p,\mathsf{m}^d}$ *is* $\tilde{L}$*-Lipschitz continuous with* $\tilde{L} = \max(L_{\mathsf{f}}, L_{\mathsf{r}})$*. It is maximally monotone if* $f(\cdot,\xi)$ *is convex for* m^p*-almost every* $\xi \in \Xi$ *and* $g(\cdot,\zeta)$ *is convex for* m^d*-almost every* $\zeta \in \mathcal{Z}$*.*

(iii) *The operator* $\mathbf{T}_{\mathsf{m}^p,\mathsf{m}^d}$ *is maximally* $\tilde{\mu}$*-strongly monotone with* $\tilde{\mu} = \min(\mu_p, \mu_d)$*.*

(iv) *For any* $x, x' \in \mathcal{H}$ *and* $y, y' \in \mathcal{K}$

$$\sup_{(x'',y'')\in\mathcal{H}\times\mathcal{K}} \|\mathbf{B}_{\mathsf{m}_x,\mathsf{m}_y}(x'',y'') - \mathbf{B}_{\mathsf{m}_{x'},\mathsf{m}_{y'}}(x'',y'')\|_{\mathcal{H}\times\mathcal{K}} \le \tilde{\beta}\tau\|(x,y)-(x',y')\|_{\mathcal{H}\times\mathcal{K}},$$

where $\tilde{\beta} = \max(\beta_{\mathsf{f}}, \beta_{\mathsf{r}})$*.*

Proof. Thanks to (64) and assumption (c), (i) follows from [23, Theorem 21.2 and Proposition 20.23]. The operator $\mathbf{B}_{\mathsf{m}^p,\mathsf{m}^d}$ is the gradient of the separable function $\Phi(x,y) = \mathsf{f}_{\mathsf{m}^p}(x) + \mathsf{r}_{\mathsf{m}^d}(y)$. From assumptions (a) and (b), we have $\nabla\mathsf{f}_{\mathsf{m}^p} = \mathbb{E}_{\xi\sim\mathsf{m}^p}(\nabla f(\cdot,\xi))$ and similarly for $\nabla\mathsf{r}_{\mathsf{m}^d}$. Thus $\Phi \in C^{1,1}_{\max(L_{\mathsf{f}},L_{\mathsf{r}})}(\mathcal{H}\times\mathcal{K})$ whence we get claim the first part of (ii). Moreover, under convexity, $\mathsf{f}_{\mathsf{m}^p}$ and $\mathsf{r}_{\mathsf{m}^d}$ are also convex, hence their gradients are maximally monotone, and we conclude using [23, Proposition 20.23]. To show (iii), observe that $\mathbf{T}_{\mathsf{m}^p,\mathsf{m}^d}(x,y) = (\partial(\mathsf{f}_{\mathsf{m}^p}+\mathsf{g})(x), \partial(\mathsf{r}_{\mathsf{m}^d}+\mathsf{h})(y)) + \mathbf{L}(x,y)$. The first part is maximally strongly monotone by (d) and [23, Example 22.4, Theorem 21.2 and Proposition 20.23]. In addition, $\mathbf{L}$ is a skew-symmetric linear operator, hence maximally monotone (see [23, Example 20.35]). We conclude using [26, Lemma 2.4]. The proof of (iv) is similar to that of Corollary 1. Indeed, let us fix $(x'',y'') \in \mathcal{H}\times\mathcal{K}$. We then have, for any $x, y \in \mathcal{H}$ and $y, y' \in \mathcal{K}$

$$\begin{aligned}
&\|\mathbf{B}_{\mathsf{m}_x,\mathsf{m}_y}(x'',y'') - \mathbf{B}_{\mathsf{m}_{x'},\mathsf{m}_{y'}}(x'',y'')\|_{\mathcal{H}\times\mathcal{K}}\\
&= \sqrt{\left\|\nabla\mathsf{f}_{\mathsf{m}_x}(x'') - \nabla\mathsf{f}_{\mathsf{m}_{x'}}(x'')\right\|^2_{\mathcal{H}} + \left\|\nabla\mathsf{r}_{\mathsf{m}_y}(y'') - \nabla\mathsf{f}_{\mathsf{m}_{y'}}(y'')\right\|^2_{\mathcal{K}}}\\
&\le \sqrt{\beta_{\mathsf{f}}^2\mathbb{W}_1(\mathsf{m}_x,\mathsf{m}_{x'})^2 + \beta_{\mathsf{r}}^2\mathbb{W}_1(\mathsf{m}_y,\mathsf{m}_{y'})^2}\\
&\le \max(\beta_{\mathsf{f}},\beta_{\mathsf{r}})\mathbb{W}_1(\mathsf{m}_x\otimes\mathsf{m}_{x'},\mathsf{m}_y\otimes\mathsf{m}_{y'})\\
&\le \max(\beta_{\mathsf{f}},\beta_{\mathsf{r}})\tau\|(x,y)-(x',y')\|_{\mathcal{H}\times\mathcal{K}}.
\end{aligned}$$

■

6.2. **Existence and uniqueness of equilibrium.** We are now in a position to prove the existence and uniqueness of an equilibrium to the problem (61)

Theorem 4 (Existence and uniqueness of equilibrium point)**.** *Under Assumption 8, the map*

$$S : (x, y) \in \mathcal{H} \mapsto \mathrm{zer}(\mathbf{T}_{\mathsf{m}_x,\mathsf{m}_y}) = \{(u, v) \in \mathcal{H} \times \mathcal{K} :\ (0, 0)_{\mathcal{H}\times\mathcal{K}} \in \mathbf{T}_{\mathsf{m}_x,\mathsf{m}_y}(u, v)\}$$

is $\tilde{\rho}$-Lipschitz with

$$\tilde{\rho} := \frac{\tau\tilde{\beta}}{\tilde{\mu}}, \tag{65}$$

where we recall that $\tilde{\mu} = \min(\mu_p, \mu_d)$ and $\tilde{\beta} = \max(\beta_{\mathsf{f}}, \beta_{\mathsf{r}})$. In particular, if $\tilde{\rho} < 1$, there is a unique equilibrium point $(\bar{x}, \bar{y})$.

An immediate consequence of this result is that (62) has a unique solution which is $(\bar{x}, \bar{y})$.

Proof. Following the same lines as in Theorem 1 and using Lemma 6(iii), we conclude. ■

Remark 6.2. Theorem 4 is to be compared to [52, Theorem 2.6]. Notice that the monotone inclusion (63) defining the equilibrium in [52], handled through a a variational inequality there to the presence of constraints on x and y, can be cast in our setting through normal cones absorbed in $\partial\mathsf{g}$ and $\partial\mathsf{h}$ under appropriate qualification conditions.

Following the reasoning of Sections 3 and 4, let us define the following gap function

$$\mathbf{E}_{\bar{x},\bar{y}}(x, y) = \mathbf{B}_{\mathsf{m}_x,\mathsf{m}_y}(x, y) - \mathbf{B}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x, y). \tag{66}$$

Mimicking the proof of Lemma 4 using Assumption 8, we prove the following.

Lemma 7. *Under Assumption 8, $\mathbf{E}_{\bar{x},\bar{y}}(.)$ is $(2\max(L_{\mathsf{f}}, L_{\mathsf{r}}) + \tau\max(\beta_{\mathsf{f}}, \beta_{\mathsf{r}}))$-Lipschitz continuous.*

6.3. **Related first and second-order dynamics.** To lighten the notation, we set $Z(t) := (x(t), y(t))$, $\bar{Z} := (\bar{x}, \bar{y})$ is the equilibrium of (61) (see Theorem 4), and $\mathsf{m}_{\bar{Z}} := \mathsf{m}_{\bar{x}} \otimes \mathsf{m}_{\bar{y}}$.

6.3.1. *First order system.* Given an initial data $Z(t_0) = (x(t_0), y(t_0)) \in \mathrm{dom}(g) \times \mathrm{dom}(h)$, we consider the following first-order system associated to the monotone inclusion (63):

$$\dot{Z}(t) + \mathbf{T}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(Z(t)) + \mathbf{E}_{\bar{Z}}(Z(t)) \ni 0_{\mathcal{H}\times\mathcal{K}}. \tag{SPDS}$$

Arguing as in Proposition 1[2] and Theorem 2 we have the following result.

Proposition 7. *Assume that Assumption 8 holds and $\tilde{\rho} < 1$. Then, for any initial data $Z(t_0) = (x(t_0), y(t_0)) \in \mathrm{dom}(g) \times \mathrm{dom}(h)$, (SPDS) admits a unique strong solution $Z : t \in [t_0, +\infty[\mapsto (x(t), y(t))$. Moreover,*

$$\|Z(t) - \bar{Z}\|_{\mathcal{H}\times\mathcal{K}} \le Ce^{-2\tilde{\mu}(1-\tilde{\rho})t},\ \forall t \ge t_0,$$

with $C = \|Z_0 - \bar{Z}\|_{\mathcal{H}\times\mathcal{K}} e^{2\tilde{\mu}(1-\tilde{\rho})t_0}$, where $\tilde{\rho}$ and $\tilde{\mu}$ are defined in (65).

[2] Here, we invoke [26, Proposition 3.12] instead of [26, Proposition 3.13] as $\mathbf{T}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}$ is a subdifferential operator perturbed by a bounded linear operator. This explains why we do not need any domain interiority assumption in this case.

6.3.2. *Second-order system.* As we have done in Section 4, here, we restrict ourselves to the smooth setting where both g and h are smooth with Lipschitz continuous gradient. In this case, the operator $\mathbf{A}$, hence $\mathbf{T}_{\mathsf{m}_{\bar{Z}}}$, is single-valued, Lipschitz continuous and strongly monotone. In explicit forms, for $Z = (x, y)$, we have

$$\mathbf{T}_{\mathsf{m}_{\bar{Z}}}(Z) = \left(\nabla_x \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x, y), -\nabla_y \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x, y)\right). \tag{67}$$

We propose the following inertial system associated to (63)

$$\ddot{Z}(t) + 2\sqrt{\tilde{\mu}}\dot{Z}(t) + \begin{pmatrix} \nabla_x \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}\left(x(t), y(t) + \frac{1}{\sqrt{\tilde{\mu}}}\dot{y}(t)\right) \\ -\nabla_y \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}\left(x(t) + \frac{1}{\sqrt{\tilde{\mu}}}\dot{x}(t), y(t)\right) \end{pmatrix} + \mathbf{E}_{\bar{Z}}(Z(t)) = 0, \tag{ISPDS$_{2\sqrt{\tilde{\mu}}}$}$$

where $\mathbf{E}_{\bar{Z}}(Z(t)) = (\mathbf{e}_{\bar{x}}(x(t)), \mathbf{e}_{\bar{y}}(y(t)))$. Unperturbed second order primal-dual dynamical systems for solving saddle point problems have been actively studied in recent years, essentially with vanishing viscous damping; see e.g., [8] and [35] and references therein. Unlike (ISEHD$_{2\sqrt{\mu},\omega}$), we did not include geometric damping in (ISPDS$_{2\sqrt{\tilde{\mu}}}$) for the sake of simplicity. Moreover, one can see that the system (ISPDS$_{2\sqrt{\tilde{\mu}}}$) is driven by the gradient of the Lagrangian but evaluated at extrapolated points, which is is standard approach for primal-dual second-order systems. In fact, one has to keep in mind that unlike (ISEHD$_{2\sqrt{\mu},\omega}$), the system (ISPDS$_{2\sqrt{\tilde{\mu}}}$) is driven by an operator deriving from the Lagrangian which is convex-concave and contains a bilinear function. This is the reason underlying the inclusion of the extrapolation step. This will also necessitate to modify the Lyapunov analysis in the proof Theorem 3 though the main ingredients will remain essentially the same.

We first state that (ISPDS$_{2\sqrt{\tilde{\mu}}}$) is well posed. This follows from the same arguments as [8, Theorem 5], using the Cauchy-Lipschitz theorem since $\nabla_x \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}$ and $\nabla_y \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}$ are globally Lipschitz continuous by assumption, and so is $\mathbf{E}_{\bar{Z}}$ by Lemma 6(iv).

Proposition 8. *Suppose that Assumption 8(a), (c) and (e) hold and, moreover, that* g *and* h *are also continuously differentiable with Lipschitz continuous gradient. Then, for any given initial condition* $(x(t_0), \dot{x}(t_0)) = (x_0, u_0) \in \mathcal{H} \times \mathcal{K}$ *and* $(y(t_0), \dot{y}(t_0)) = (y_0, v_0) \in \mathcal{H} \times \mathcal{K}$ *the evolution system* (ISPDS$_{2\sqrt{\tilde{\mu}}}$) *has a unique strong global solution* $Z(\cdot) = (x(\cdot), y(\cdot))$ *with*

- $Z \in C^1([0, +\infty[; \mathcal{H} \times \mathcal{K})$;
- Z *and* $\dot{Z}$ *are absolutely continuous on every compact subset of the interior of* $[t_0, +\infty[$ *(hence almost everywhere differentiable);*
- *for almost all* $t \in [t_0, +\infty[$, (ISPDS$_{2\sqrt{\tilde{\mu}}}$) *holds with* $Z(t_0) = (x_0, y_0)$ *and* $\dot{Z}(t_0) = (u_0, v_0)$.

Let us now move to the convergence properties of (ISPDS$_{2\sqrt{\tilde{\mu}}}$). We define the following energy function $\mathcal{V} : [t_0, \infty[\to \mathbb{R}^+$ by

$$\mathcal{V}(t) := \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x(t), \bar{y}) - \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(\bar{x}, y(t)) + \frac{1}{2}\|v(t)\|^2_{\mathcal{H}\times\mathcal{K}}$$

with

$$v(t) := \sqrt{\tilde{\mu}}\left(Z(t) - \bar{Z}\right) + \dot{Z}(t),$$

Observe that the Lagrangian gap in $\mathcal{V}$ is nonnegative and convex in $Z(t)$.

Theorem 5. *Assume that Assumption 8 holds and, moreover, that* g *and* h *are also continuously differentiable with Lipschitz continuous gradient. Let* $t \in [t_0, \infty[\mapsto (x(t), y(y))$ *be the solution of* (ISPDS$_{2\sqrt{\tilde{\mu}}}$). *Suppose that* $\tilde{\rho} < \frac{\sqrt{2}}{4}$, *where* $\tilde{\rho}$ *is as given in* (65). *We then have for all* $t \geq t_0$*:*

$$\frac{\tilde{\mu}}{2}\|Z(t) - \bar{Z}\|^2_{\mathcal{H}\times\mathcal{K}} \leq \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x(t), \bar{y}) - \mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(\bar{x}, y(t)) \leq \mathcal{V}(t_0)e^{-\frac{\sqrt{\tilde{\mu}}}{4}(t-t_0)}. \tag{68}$$

Proof. To lighten notation, we drop dependence of the norm and inner product on the underlying space which is to be understood from the context. At first, we have, using ($\mathrm{ISPDS}_{2\sqrt{\tilde\mu}}$)

$$\begin{aligned}
\dot v(t) = \sqrt{\tilde\mu}\dot Z(t) + \ddot Z(t) &= -\sqrt{\tilde\mu}\dot Z(t) - \begin{pmatrix} \nabla_x \mathcal{L}_{\mathsf{m}_{\bar x},\mathsf{m}_{\bar y}}\left(x(t), y(t)+\frac{1}{\sqrt{\tilde\mu}}\dot y(t)\right) \\ -\nabla_y \mathcal{L}_{\mathsf{m}_{\bar x},\mathsf{m}_{\bar y}}\left(x(t)+\frac{1}{\sqrt{\tilde\mu}}\dot x(t), y(t)\right)\end{pmatrix} - \mathbf{E}_{\bar Z}(Z(t)) \\
&= -\sqrt{\tilde\mu}\dot Z(t) - \begin{pmatrix} \nabla(\mathsf{f}_{\mathsf{m}_{\bar x}} + \mathsf{g})(x(t)) \\ \nabla(\mathsf{r}_{\mathsf{m}_{\bar y}} + \mathsf{h})(y(t)) \end{pmatrix} - \mathbf{L}\left(Z(t) + \frac{1}{\sqrt{\tilde\mu}}\dot Z(t)\right) - \mathbf{E}_{\bar Z}(Z(t)) \\
&= -\sqrt{\tilde\mu}\dot Z(t) - \mathbf{T}_{\mathsf{m}_{\bar Z}}(Z(t)) - \frac{1}{\sqrt{\tilde\mu}}\mathbf{L}\dot Z(t) - \mathbf{E}_{\bar Z}(Z(t)).
\end{aligned}$$

We thus get,

$$\begin{aligned}
\dot{\mathcal{V}}(t) &= \left\langle \left(\nabla_x \mathcal{L}_{\mathsf{m}_{\bar x},\mathsf{m}_{\bar y}}(x(t),\bar y), -\nabla_y \mathcal{L}_{\mathsf{m}_{\bar x},\mathsf{m}_{\bar y}}(\bar x, y(t))\right), \dot Z(t)\right\rangle + \langle v(t), \dot v(t)\rangle \\
&= \left\langle \left(\nabla(\mathsf{f}_{\mathsf{m}_{\bar x}} + \mathsf{g})(x(t)), \nabla(\mathsf{r}_{\mathsf{m}_{\bar y}} + \mathsf{h})(y(t))\right) + \mathbf{L}\bar Z, \dot Z(t)\right\rangle + \langle v(t), \dot v(t)\rangle \\
&= \langle \mathbf{T}_{\mathsf{m}_{\bar Z}}(Z(t)) + \mathbf{L}(\bar Z - Z(t)), \dot Z(t)\rangle \\
&\quad + \langle \sqrt{\tilde\mu}\left(Z(t) - \bar Z\right) + \dot Z(t), -\sqrt{\tilde\mu}\dot Z(t) - \mathbf{T}_{\mathsf{m}_{\bar Z}}(Z(t)) - \frac{1}{\sqrt{\tilde\mu}}\mathbf{L}\dot Z(t) - \mathbf{E}_{\bar Z}(Z(t))\rangle.
\end{aligned}$$

After some simplifications and using that $\mathbf{L}$ is skew-symmetric, we arrive at

$$\dot{\mathcal{V}}(t) + \sqrt{\tilde\mu}\langle \mathbf{T}_{\mathsf{m}_{\bar Z}}(Z(t)), Z(t) - \bar Z\rangle + \tilde\mu\langle \dot Z(t), Z(t) - \bar Z\rangle + \sqrt{\tilde\mu}\|\dot Z(t)\|^2 = -\langle v(t), \mathbf{E}_{\bar Z}(Z(t))\rangle. \quad (69)$$

Rather than using $\tilde\mu$-strong monotonicity of $\mathbf{T}_{\mathsf{m}_{\bar Z}}$ (Lemma 6(iii)), we will now prove an alternative bound on the first inner product in (69) that will be useful for our Lyapunov analysis. We have

$$\begin{aligned}
&\langle \mathbf{T}_{\mathsf{m}_{\bar Z}}(Z(t)), Z(t) - \bar Z\rangle \\
&= \langle \nabla(\mathsf{f}_{\mathsf{m}_{\bar x}} + \mathsf{g})(x(t)), x(t) - \bar x\rangle + \langle \nabla(\mathsf{r}_{\mathsf{m}_{\bar y}} + \mathsf{h})(y(t)), y(t) - \bar y\rangle + \langle \mathbf{L}Z(t), Z(t) - \bar Z\rangle \\
&= \langle \nabla(\mathsf{f}_{\mathsf{m}_{\bar x}} + \mathsf{g})(x(t)), x(t) - \bar x\rangle + \langle \nabla(\mathsf{r}_{\mathsf{m}_{\bar y}} + \mathsf{h})(y(t)), y(t) - \bar y\rangle - \langle \mathrm{K}^* y(t), \bar x\rangle + \langle \mathrm{K}x(t), \bar y\rangle.
\end{aligned}$$

Now, by Assumption 8(d), we have

$$\begin{aligned}
\langle \nabla(\mathsf{f}_{\mathsf{m}_{\bar x}} + \mathsf{g})(x(t)), x(t) - \bar x\rangle &\geq (\mathsf{f}_{\mathsf{m}_{\bar x}} + \mathsf{g})(x(t)) - (\mathsf{f}_{\mathsf{m}_{\bar x}} + \mathsf{g})(\bar x) + \frac{\mu_p}{2}\|x(t) - \bar x\|^2 \\
\langle \nabla(\mathsf{r}_{\mathsf{m}_{\bar y}} + \mathsf{h})(y(t)), y(t) - \bar y\rangle &\geq (\mathsf{r}_{\mathsf{m}_{\bar y}} + \mathsf{h})(y(t)) - (\mathsf{r}_{\mathsf{m}_{\bar y}} + \mathsf{h})(\bar y) + \frac{\mu_d}{2}\|y(t) - \bar y\|^2.
\end{aligned}$$

Summing we get that

$$\begin{aligned}
&\langle \mathbf{T}_{\mathsf{m}_{\bar Z}}(Z(t)), Z(t) - \bar Z\rangle \\
&\geq (\mathsf{f}_{\mathsf{m}_{\bar x}} + \mathsf{g})(x(t)) - (\mathsf{r}_{\mathsf{m}_{\bar y}} + \mathsf{h})(\bar y) + (\mathsf{r}_{\mathsf{m}_{\bar y}} + \mathsf{h})(y(t)) - (\mathsf{f}_{\mathsf{m}_{\bar x}} + \mathsf{g})(\bar x) - \langle \mathrm{K}^* y(t), \bar x\rangle + \langle \mathrm{K}x(t), \bar y\rangle + \frac{\tilde\mu}{2}\|Z(t) - \bar Z\|^2 \\
&= \mathcal{L}_{\mathsf{m}_{\bar x},\mathsf{m}_{\bar y}}(x(t), \bar y) - \mathcal{L}_{\mathsf{m}_{\bar x},\mathsf{m}_{\bar y}}(\bar x, y(t)) + \frac{\tilde\mu}{2}\|Z(t) - \bar Z\|^2.
\end{aligned}$$

Plugging this into (69), we get

$$\begin{aligned}&\dot{\mathcal{V}}(t)+\sqrt{\tilde{\mu}}\left(\mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x(t),\bar{y})-\mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(\bar{x},y(t))+\frac{\tilde{\mu}}{2}\|Z(t)-\bar{Z}\|^2+\sqrt{\tilde{\mu}}\langle\dot{Z}(t),Z(t)-\bar{Z}\rangle+\|\dot{Z}(t)\|^2\right)\\&=\dot{\mathcal{V}}(t)+\sqrt{\tilde{\mu}}\left(\mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x(t),\bar{y})-\mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(\bar{x},y(t))+\frac{1}{2}\|\sqrt{\tilde{\mu}}(Z(t)-\bar{Z})+\dot{Z}(t)\|^2\right)+\frac{\tilde{\mu}}{2}\|\dot{Z}(t)\|^2\\&=\dot{\mathcal{V}}(t)+\sqrt{\tilde{\mu}}\mathcal{V}(t)+\frac{\sqrt{\tilde{\mu}}}{2}\|\dot{Z}(t)\|^2\leq\|v(t)\|\|\mathbf{E}_{\bar{Z}}(Z(t))\|.\end{aligned}\tag{70}$$

where we used Cauchy-Schwarz inequality in the last line. Arguing as above, using again Assumption 8(d), we have

$$\mathcal{V}(t)\geq\mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(x(t),\bar{y})-\mathcal{L}_{\mathsf{m}_{\bar{x}},\mathsf{m}_{\bar{y}}}(\bar{x},y(t))\geq\frac{\tilde{\mu}}{2}\|Z(t)-\bar{Z}\|^2.\tag{71}$$

Using this in (70) yields the bound

$$\dot{\mathcal{V}}(t)+\frac{\sqrt{\tilde{\mu}}}{2}\mathcal{V}(t)+\frac{\sqrt{\tilde{\mu}}}{2}\|\dot{Z}(t)\|^2+\frac{\tilde{\mu}^{3/2}}{4}\|Z(t)-\bar{Z}\|^2\leq\|v(t)\|\|\mathbf{E}_{\bar{Z}}(Z(t))\|.\tag{72}$$

We have, using Young's inequality, Lemma 6(iv) and the fact that $\mathcal{V}(t)\geq\frac{1}{2}\|v(t)\|^2$

$$\|v(t)\|\|\mathbf{E}_{\bar{Z}}(Z(t))\|\leq\frac{\sqrt{\tilde{\mu}}}{8}\|v(t)\|^2+\frac{2}{\sqrt{\tilde{\mu}}}\|\mathbf{E}_{\bar{Z}}(Z(t))\|^2\leq\frac{\sqrt{\tilde{\mu}}}{4}\mathcal{V}(t)+\frac{2\tilde{\beta}^2\tau^2}{\sqrt{\tilde{\mu}}}\|Z(t)-\bar{Z}\|^2.$$

Inserting into (72) and rearranging gives

$$\dot{\mathcal{V}}(t)+\frac{\sqrt{\tilde{\mu}}}{4}\mathcal{V}(t)+\frac{\sqrt{\tilde{\mu}}}{2}\|\dot{Z}(t)\|^2+\sqrt{\tilde{\mu}}\left(\frac{\tilde{\mu}}{4}-\frac{2\tilde{\beta}^2\tau^2}{\tilde{\mu}}\right)\|Z(t)-\bar{Z}\|^2\leq 0.$$

Since $\frac{\tilde{\beta}\tau}{\tilde{\mu}}<\frac{\sqrt{2}}{4}$ by assumption, we have

$$\dot{\mathcal{V}}(t)+\frac{\sqrt{\tilde{\mu}}}{4}\mathcal{V}(t)\leq 0,$$

which gives after integration yields

$$\mathcal{V}(t)\leq\mathcal{V}(t_0)e^{-\frac{\sqrt{\tilde{\mu}}}{4}(t-t_0)},\ \forall t\geq t_0.\tag{73}$$

Plugging this into (71), we prove the claim. ■

7. Comments, extensions and future work

In this paper, we adopted a dynamical system approach to study some stochastic optimization problems with state-dependent distributions. We investigated the existence and uniqueness of equilibrium points, well-posedness as well as convergence properties of the trajectories, for both first and second-order dynamics. We highlighted some dynamical and geometrical properties of the state-dependent distributions suggesting that the natural framework to study problems of the form (5) is the one of *metric random walk spaces*. More particularly, the notion of coarse Ricci curvature gives a new insight on the geometrical hidden structure of this kind of problems. Finally, we discussed as an application the inertial primal-dual algorithm. We present here some ongoing works, possible extensions as well as some open problems.

Inertial algorithms. Relying on the discretization of the dynamics studied in Section 4 and Section 6, more specifically, ($\mathrm{ISEHD}_{2\sqrt{\mu},\omega}$) and ($\mathrm{ISPDS}_{2\sqrt{\mu}}$) , we obtain new inertial algorithms with Hessian-driven damping for stochastic optimization problems with decision-dependent distributions. These algorithms exhibit rapid convergence properties and can also be adapted to the nonsmooth case. This is being addressed in ongoing work.

Implicit Hessian damping. We focused in Section 4 on the explicit Hessian damping in the smooth case. Yet, it is possible to consider implicit damping as in [9, 14]

$$\ddot{x}(t)+\gamma(t)\dot{x}(t)+\nabla G_{\mathsf{m}_{\bar{x}}}\left(x(t)+\omega\dot{x}(t)\right)+\mathbf{e}_{\bar{x}}(x(t))=0, \qquad (\mathrm{ISIHD}_{\mathsf{m},\gamma})$$

The dynamics ($\mathrm{ISIHD}_{\mathsf{m},\gamma}$) is referred to as an Inertial System with Implicit Hessian damping, since one can observe, using Taylor expansion

$$\nabla G_{\mathsf{m}_{\bar{x}}}\left(x(t)+\omega\dot{x}(t)\right)\approx\nabla G_{\mathsf{m}_{\bar{x}}}x(t)+\nabla^2 G_{\mathsf{m}_{\bar{x}}}(x(t))\dot{x}(t).$$

As it was observed in [14], higher-order moments of the perturbation $\mathbf{e}_{\bar{x}}$ are required to get fast convergence guarantees in the implicit case compared to the explicit one. Since in our analysis (see Theorem 3) no integrability assumption on $\mathbf{e}_{\bar{x}}$ is needed, it is interesting to investigate the effect of implicit Hessian damping, both in the smooth and nonsmooth cases.

Tikhonov-regularization. In Section 4 we restricted ourselves, for sake of simplicity, the analysis to the case where the operator $F_{\mathsf{m}_{\bar{x}}}$ is smooth. However, it is possible to consider second-order dynamics for general (and possibly nonpotential) operators, by considering, for $\lambda>0$ the following dynamic

$$\ddot{x}(t)+\gamma(t)\dot{x}(t)+F^{\lambda(t)}_{\mathsf{m}_{\bar{x}(t)}}(x(t))+\omega\frac{\mathrm{d}}{\mathrm{d}t}\left(F^{\lambda(t)}_{\mathsf{m}_{\bar{x}(t)}}(x(t))\right)+\mathbf{e}_{\bar{x}}(x(t))=0. \qquad (\mathrm{ISEHD}_{\lambda,\gamma})$$

where $F^{\lambda}_{\mathsf{m}_{\bar{x}}}$ is the so-called Yosida approximation of $F_{\mathsf{m}_{\bar{x}}}$ defined by $F^{\lambda}_{\mathsf{m}_{\bar{x}}}=\frac{1}{\lambda}\left(\mathrm{id}-J_{\lambda F_{\mathsf{m}_{\bar{x}}}}\right)$ and $J_{\lambda F_{\mathsf{m}_{\bar{x}}}}=(\mathrm{id}+JF_{\mathsf{m}_{\bar{x}}})^{-1}$ is the resolvent of $F_{\mathsf{m}_{\bar{x}}}$. This approach comes with several advantages. First, the Yosida approximation is single-valued so that there is no nonsmoothness to take care of. In addition one can exploit the $\lambda-$cocoercivity of $F^{\lambda}_{\mathsf{m}_{\bar{x}}}$ and the fact that $\operatorname{zer} F_{\mathsf{m}_{\bar{x}}}=\operatorname{zer} F^{\lambda}_{\mathsf{m}_{\bar{x}}}$. The approach was used in [17] for $\omega=0$ and in the recent work [12] for Newton-like dynamics. We are exploring the adaptation of this techniques to stochastic monotone inclusions with state-dependent distribution in an ongoing work.

Weaker Assumptions. We have seen that one of the crucial assumptions in the analysis is Assumption 2, which concerns the Lipschitz behavior of the distribution $(\mathsf{m}_x)_x$. A natural question that arises is what happens under a weaker assumption. For example, when $x\mapsto\mathsf{m}_x$ is Hölder continuous. We are not aware of any existing results in this direction. We plan to investigate this in future work.

Acknowledgement

The work of H.E was supported by the ANR grant, reference: ANR-20-CE92-0037.

A. Gronwall inequalities

In this section we list several auxiliary results that we make use of in the paper.

Lemma 8 (Gronwall's lemma: differential form)**.** *Let u,v be two C^0 (resp. C^1) nonnegative function on $[0,T]$ and let w be a continuous function on $[0,T]$. We assume that*

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}u^2(t)\leq w(t)u^2(t)+u(t)v(t)\ \textit{on}\ (0,T), \qquad (74)$$

then, for any $t \in [0, T]$

$$u(t) \leq u(0)e^{K(t)} + \int_0^t v(s)e^{K(t)-K(s)}\mathrm{d}s, \tag{75}$$

where $K(t) = \int_0^t w(s)\mathrm{d}s$*.*

Lemma 9. *[26, Lemma A.5] Let* $v \in L^1(t_0, T; \mathbb{R}^+)$ *and* $u \in C^0(t_0, T)$ *such that*

$$\frac{1}{2}u^2(t) \leq \frac{1}{2}c^2 + \int_{t_0}^t u(s)v(s)\mathrm{d}s,$$

for some $c \geq 0$ *for all* $t \in [t_0, T]$*. Then*

$$|u(t)| \leq c + \int_{t_0}^t v(s)\mathrm{d}s.$$

B. Banach Fixed point theorem & Picard iterative method

Theorem 6 (see, e.g., [6, 27])**.** *Let* $(\mathcal{X}, \mathrm{d})$ *be a complete metric space and* $S : \mathcal{X} \to \mathcal{X}$ *be a strict contraction, i.e., there exists a constant* $\varrho < 1$ *such that*

$$\mathrm{d}(S(x), S(y)) \leq \varrho \mathrm{d}(x, y), \forall x, y \in \mathcal{X}.$$

Then, there exists a unique $\bar{x} \in \mathcal{X}$ *such that* $S(\bar{x}) = \bar{x}$*. Moreover, for any* $x_0 \in \mathcal{X}$*, the sequence starting from* x_0 *with* $x_{n+1} = S(x_n)$ *for all* $n \in \mathbb{N}$ *converges to* $\bar{x}$ *as* n *goes to* ∞*.*

References

[1] F. Alvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. *SIAM J. Control Optim.*, 38(4):1102–1119, 2000.

[2] F. Alvarez and H. Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Anal.*, 9(1-2):3–11, 2001.

[3] F. Alvarez, H. Attouch, J. Bolte, and P. Redont. A second-order gradient-like dissipative dynamical system with hessian-driven damping.: Application to optimization and mechanics. *Journal de mathématiques pures et appliquées*, 81(8):747–779, 2002.

[4] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures.* Basel: Birkhäuser, 2nd ed. edition, 2008.

[5] V. Apidopoulos, J.-F. Aujol, and C. Dossal. Convergence rate of inertial forward–backward algorithm beyond Nesterov's rule. *Mathematical Programming*, 180(1):137–156, 2020.

[6] H. Attouch, G. Buttazzo, and G. Michaille. *Variational analysis in Sobolev and BV spaces*, volume 17 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, second edition, 2014. Applications to PDEs and optimization.

[7] H. Attouch, A. Cabot, and P. Redont. The dynamics of elastic shocks via epigraphical regularization of a differential inclusion. Barrier and penalty approximations. *Adv. Math. Sci. Appl.*, 12(1):273–306, 2002.

[8] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. Fast convergence of dynamical ADMM via time scaling of damped inertial dynamics. *J. Optim. Theory Appl.*, 193(1-3):704–736, 2022.

[9] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. First-order optimization algorithms via inertial systems with Hessian driven damping. *Math. Program.*, 193(1, Ser. A):113–155, 2022.

[10] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168:123–175, 2018.

[11] H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:2, 2019.

[12] H. Attouch and S. Csaba László. Continuous Newton-like inertial dynamics for monotone inclusions. *Set-Valued Var. Anal.*, 29(3):555–581, 2021.

[13] H. Attouch and A. Damlamian. On multivalued evolution equations in hilbert spaces. *Israel Journal of Mathematics*, 12:373–390, 1972.

[14] H. Attouch, J. Fadili, and V. Kungurtsev. On the effect of perturbations in first-order optimization methods with inertia and Hessian driven damping. *Evol. Equ. Control Theory*, 12(1):71–117, 2023.

[15] H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method, i. the continuous dynaamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(01):1–34, 2000.

[16] H. Attouch and J. Peypouquet. The rate of convergence of nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

[17] H. Attouch and J. Peypouquet. Convergence of inertial dynamics and proximal algorithms governed by maximally monotone operators. *Math. Program.*, 174(1-2 (B)):391–432, 2019.

[18] H. Attouch, J. Peypouquet, and P. Redont. A dynamical approach to an inertial forward-backward algorithm for convex minimization. *SIAM Journal on Optimization*, 24(1):232–256, 2014.

[19] H. Attouch, J. Peypouquet, and P. Redont. Fast convex optimization via inertial dynamics with hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783, 2016.

[20] H. Bahouri, J.-Y. Chemin, and R. Danchin. *Fourier analysis and nonlinear partial differential equations*, volume 343 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Heidelberg, 2011.

[21] J. Baillon and H. Brézis. Une remarque sur le comportement asymptotique des semigroupes non linéaires. *Houston J. Math.*, 2:5–7, 1976.

[22] V. Barbu. *Nonlinear differential equations of monotone types in Banach spaces*. Springer Monographs in Mathematics. Springer, New York, 2010.

[23] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. With a foreword by Hédy Attouch.

[24] P. Bianchi, W. Hachem, and A. Salim. A fully stochastic primal-dual algorithm. *Optimization Letters*, 15(2):701–710, 2021.

[25] R. I. Boţ, E. R. Csetnek, and S. C. László. Tikhonov regularization of a second order dynamical system with Hessian driven damping. *Math. Program.*, 189(1-2 (B)):151–186, 2021.

[26] H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland Mathematics Studies, No. 5. North-Holland Publishing Co., Amsterdam-London; American Elsevier Publishing Co., Inc., New York, 1973.

[27] H. Brézis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[28] R. E. Bruck Jr. Asymptotic convergence of nonlinear contraction semigroups in hilbert space. *Journal of Functional Analysis*, 18(1):15–26, 1975.

[29] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40:120–145, 2011.

[30] L. Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of optimization theory and applications*, 158(2):460–479, 2013.

[31] J. Cutler, M. Díaz, and D. Drusvyatskiy. Stochastic approximation with decision-dependent distributions: asymptotic normality and optimality. *arXiv preprint arXiv:2207.04173*, 2022.

[32] A. Derrow-Pinion, J. She, D. Wong, O. Lange, T. Hester, L. Perez, M. Nunkesser, S. Lee, X. Guo, B. Wiltshire, P. W. Battaglia, V. Gupta, A. Li, Z. Xu, A. Sanchez-Gonzalez, Y. Li, and P. Velickovic. Eta prediction with graph neural networks in google maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pages 3767–3776, New York, NY, USA, 2021. Association for Computing Machinery.

[33] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018.

[34] D. Drusvyatskiy and L. Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

[35] X. He, R. Hu, and Y. Fang. A second order primalâ-dual dynamical system for a convex-concave bilinear saddle point problem. *Applied Mathematics & Optimization*, 89(2):30, 2024.

[36] O. Hernández-Lerma and J. B. Lasserre. *Markov chains and invariant probabilities*, volume 211 of *Prog. Math.* Basel: Birkhäuser, 2003.

[37] D. Kim. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190(1):57–87, 2021.

[38] S. C. László. Convergence rates for an inertial algorithm of gradient type associated to a smooth non-convex minimization. *Mathematical Programming*, 190(1):285–329, 2021.

[39] T. Lin and M. I. Jordan. A control-theoretic perspective on optimal high-order optimization. *Mathematical Programming*, pages 1–47, 2022.

[40] J. Macfarlane. Your navigation app is making traffic unmanageable. *IEEE Spectrum*, 19, 2019.

[41] J. Mazón, M. Solera-Diana, and J. Toledo-Melero. *Variational and Diffusion Problems in Random Walk Spaces*. Progress in Nonlinear Differential Equations and Their Applications. Springer Nature Switzerland, 2023.

[42] C. Mendler-Dünner, J. C. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performative prediction. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[43] Y. Nesterov. A method for solving the convex programming problem with convergence rate o(1/k^2). *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

[44] Y. Ollivier. Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.*, 256(3):810–864, 2009.

[45] Y. Ollivier. A survey of Ricci curvature for metric spaces and markov chains. In *Probabilistic approach to geometry*, volume 57, pages 343–382. Mathematical Society of Japan, 2010.

[46] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 13–18 Jul 2020.

[47] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[48] H. Raguet, J. Fadili, and G. Peyré. A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

[49] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2022.

[50] W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *J. Mach. Learn. Res.*, 17:43, 2016. Id/No 153.

[51] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

[52] K. Wood and E. Dall'Anese. Stochastic saddle point problems with decision-dependent distributions. *SIAM Journal on Optimization*, 33(3):1943–1967, 2023.