PAPER

A bilevel optimization method for inverse mean-field games*

IOP Publishing Inverse Problems
Inverse Problems **40** (2024) 105016 (43pp) https://doi.org/10.1088/1361-6420/ad75b0

A bilevel optimization method for inverse mean-field games*

Jiajia Yu[1,**], **Quan Xiao**[2], **Tianyi Chen**[2] and **Rongjie Lai**[3]

[1] Department of Mathematics, Duke University, Durham, NC, United States of America
[2] Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, United States of America
[3] Department of Mathematics, Purdue University, West Lafayette, IN, United States of America

E-mail: jiajia.yu@duke.edu, xiaoq5@rpi.edu, chent18@rpi.edu and lairj@purdue.edu

Received 10 January 2024; revised 13 August 2024
Accepted for publication 30 August 2024
Published 13 September 2024

Abstract

In this paper, we introduce a bilevel optimization framework for addressing inverse mean-field games, alongside an exploration of numerical methods tailored for this bilevel problem. The primary benefit of our bilevel formulation lies in maintaining the convexity of the objective function and the linearity of constraints in the forward problem. Our paper focuses on inverse mean-field games characterized by unknown obstacles and metrics. We show numerical stability for these two types of inverse problems. More importantly, we, for the first time, establish the identifiability of the inverse mean-field game with unknown obstacles via the solution of the resultant bilevel problem. The bilevel approach enables us to employ an alternating gradient-based optimization algorithm with a provable convergence guarantee. To validate the effectiveness of our methods in solving the inverse problems, we have designed comprehensive numerical experiments, providing empirical evidence of its efficacy.

Keywords: mean-field games, inverse problems, bilevel optimization, alternating gradient method

* This work is supported in part by NSF DMS-2134168.
** Author to whom any correspondence should be addressed.

1. Introduction

Mean-field games study the Nash Equilibrium in a non-cooperative game with infinitely many agents. In the game, each agent aims to minimize a combination of dynamic cost, interaction cost, and terminal cost by controlling its own state trajectory. At the Nash Equilibrium, the agents cannot unilaterally reduce their costs. The theory is proposed in [4, 12, 22] and has attracted increasing attention since then.

In most existing works, knowing the cost functions is required to solve mean-field games. However, in practice, these cost functions are not always easy to obtain. In contrast, the state distribution, the strategies of agents, and sometimes the value function at the Nash Equilibrium can be observed. Thus, a natural question arises: Can we learn the cost functions from the Nash Equilibrium? We refer to this as the inverse mean-field game problem, and to the original one as the forward problem.

Unlike the forward problem, relatively few studies focus on inverse mean-field games. Kachroo *et al* [15] derives two traffic flow models as the solutions of non-viscous mean-field games. Ding *et al* [8] reconstructs the underlying metric in the dynamic cost and the kernel in the non-local interaction cost from the possibly noisy observation of agent distribution and strategy. Chow *et al* [7] learns the running cost from population density and strategy on a given boundary. Guo *et al* [10] infers the full information of population density, strategy and the model from partial and noisy observation of the density and model through Gaussian Process, a Bayesian non-parametric technique for supervised learning. Klibanov *et al* [21] proposes a convexification method with global convergence for recovering the interaction coefficient function from a single measurement data. References [24–26, 29] establish the theoretical unique identifiability result for a general class of mean-field games, mean-field game boundary problems and multipopulation mean-field games, where infinite pairs of training data are required in the proof. Following [18], a series of works [13, 16, 17, 19, 20] study the stability and uniqueness of inverse mean-field game through Carleman estimates.

In this paper, we study a typical class of forward problems, the potential mean-field games. Applications like crowd motion [30] and generative models [23] have the formulations of potential mean-field games. In a potential mean-field game, the Nash Equilibrium is a pair of agent distribution ρ and strategy $\mathbf{m}$ minimizing a cost $\mathcal{L}$ which consists of the dynamic cost L, the interaction cost $\mathcal{F}_I$ and the terminal cost $\mathcal{F}_T$, under a constraint $\mathcal{C}(\mu_0)$ for density and strategy evolution dynamics:

$$(\rho^*, \mathbf{m}^*) := \underset{\rho, \mathbf{m} \in \mathcal{C}(\mu_0)}{\operatorname{argmin}} \ \mathcal{L}(\rho, \mathbf{m}; L, \mathcal{F}_I, \mathcal{F}_T). \tag{1}$$

The inverse problem is to identify $L, \mathcal{F}_I$ or $\mathcal{F}_T$ given $(\rho^*, \mathbf{m}^*)$. Typical choices of $L, \mathcal{F}_I$ and $\mathcal{F}_T$ make (1) a (strongly) convex optimization problem with linear constraint. Taking L unknown and $\mathcal{F}_I, \mathcal{F}_T$ known as an example, we thus consider the following bilevel optimization problem

$$\begin{aligned} \min_{L} \quad & \mathcal{D}\left((\rho^*, \mathbf{m}^*), (\rho(L), \mathbf{m}(L))\right) + \mathcal{R}(L) \\ \text{s.t.} \quad & (\rho(L), \mathbf{m}(L)) := \underset{\rho, \mathbf{m} \in \mathcal{C}(\mu_0)}{\operatorname{argmin}} \ \mathcal{L}(\rho, \mathbf{m}; L, \mathcal{F}_I, \mathcal{F}_T). \end{aligned} \tag{2}$$

Here $\mathcal{D}$ is a fidelity term and $\mathcal{R}$ is a regularity term. Existing works [7, 8] use the nonlinear and nonconvex PDE optimality conditions as constraints. Consequently, achieving a theoretical convergence guarantee is challenging. In contrast, we propose a bilevel formulation for inverse mean-field games, which directly incorporates the forward problem as the constraint. This bilevel formulation maintains the desirable convex-linear structure of the forward problem (the

lower-level problem) and enables us to adopt a gradient-based bilevel optimization algorithm [6, 11, 14, 31, 32] to address the inverse problem (2). Moreover, leveraging this convex-linear structure, we have developed a convergence result, demonstrating that our algorithm converges to the stationary point of the bilevel problem.

A common question in inverse mean-field games concerns the stability and unique identifiability of the unknown parameter or function relative to the data. In our setting, we ask whether it is possible to uniquely recover the cost functions from a single pair of observations $(\rho^*, \mathbf{m}^*)$ and whether the recovered cost function continuously depends on these observations. This setup differs significantly from the theoretical work discussed in [24, 25]. In those studies, the authors demonstrate that if the interaction and terminal costs are local, holomorphic in $\rho(\mathbf{x}, t)$, and meet zero admissibility conditions, then it is possible to uniquely recover them from infinitely many observations either throughout the full domain or on its boundary. However, in our case, the cost function for a crowd motion model typically does not satisfy the zero admissibility condition. Moreover, obtaining infinitely many observations is not feasible in practice. In this work, we establish stability results for a general model and unique identifiability results for crowd motion models at a discrete level. Specifically, for a general model, we prove that our model can achieve a close solution to the ground truth with noisy observation, and for the crowd motion model, we prove that only one pair of complete observation $(\rho^*, \mathbf{m}^*)$ is sufficient to uniquely recover the obstacle function, up to a constant. Thus, compared to the requirement of infinitely many observations in [24], our result is more practical and offers insights into what constitutes an effective observation for accurately recovering the ground truth.

Contribution: We summarize our contributions as follows.

1. We propose a bilevel optimization framework for modeling inverse mean-field games.
2. We study a general model of mean-field games and show that the unknown cost parameters continuously depend on the observation of the Nash Equilibrium.
3. For the crowd motion model, we prove that up to a constant, the ground truth obstacle function is the unique minimizer to the bilevel optimization problem of the inverse mean-field game.
4. We apply an alternating gradient-based bilevel optimization algorithm to solve inverse mean-field games and prove the algorithm converges to the stationary point of the bilevel problem.
5. We implement the algorithm and illustrate the effectiveness of our model and algorithm with comprehensive numerical experiments.

Organization: The paper is organized as follows. In section 2, we briefly review the potential mean-field games and provide two examples of forward mean-field game models whose inverse problem will be addressed in this paper. In section 3, we provide the bilevel optimization model for inverse mean-field games and discretize the model. We also state the stability of both models and the unique identifiability of the inverse crowd motion model and prove them in section 5. In section 4, we present the algorithm to solve our bilevel model for inverse mean-field games and prove the convergence in section 5. In section 6, we demonstrate our model and algorithm with experiments. Finally, we conclude our work in section 7.

Notation: We summarize the notations frequently used throughout this paper in table 1.

Table 1. Notations.

MFG	$\rho, \mathbf{m}, \phi$: density, momentum and value functions g, b: metric and obstacle functions $\mathcal{L}$: objective function of a forward potential MFG γ_I, γ_T: coefficient of entropy and KL divergence $\mathcal{C}(\mu_0)$: constraint set of $(\rho, \mathbf{m})$ with $\rho(\cdot, 0) = \mu_0$ $\mathcal{C}_g, \mathcal{C}_b$: constraint set of metric g and b
Bilevel optimization	η: the lower-level variable of bilevel optimization, corresponding to $(\rho, \mathbf{m})$ in inverse MFG setting ξ: the upper-level variable of bilevel optimization, corresponding to g, b or other parameters for unknown cost functions in inverse MFG setting $\mathcal{L}(\eta;\xi), \mathcal{U}(\eta;\xi)$: lower-level and upper-level cost functions $\eta^*(\xi) := \text{argmin}_{\eta \in H} \mathcal{L}(\eta;\xi)$: lower-level optimizer for given upper-level variable ξ $u(\xi) := \mathcal{U}(\eta^*(\xi), \xi)$: upper-level objective to minimizer $H = \{\eta \mid A\eta = c\}$: the constraint set of lower-level variable, corresponding to $\mathcal{C}(\mu_0)$ in inverse MFG setting Ξ: the constraint set of upper-level variable, corresponding to $\mathcal{C}_g, \mathcal{C}_b$ in inverse MFG setting
Alternating gradient	k_u, k_l: upper-level and lower-level number of iterations τ_u, τ_l: upper-level and lower-level stepsizes
Discretization	$\mathcal{G}^\rho, \mathcal{G}^{m^x}, \mathcal{G}^{m^y}, \mathcal{G}^\phi$: discrete grids where ρ, m^x, m^y and ϕ locate i_t, i_x, i_y: indices on t, x and y direction I_t, I_x, I_y: interpolation operator on t, x and y direction D_t, D_x, D_y: difference operator on t, x and y direction

2. A review of potential mean-field games

In this section, we first review potential mean-field games and their optimality conditions [4, 12, 22]. Then we present two example problems that we would like to solve in the inverse problem setup.

Consider a problem defined spatially on $\Omega \subset \mathbb{R}^d$ and temporally on $[0,1]$. $\rho : \Omega \times [0,1] \to \mathbb{R}$ is the state density. $\mathbf{v} : \Omega \times [0,1] \to \mathbb{R}^d$ represents the velocity (control) field of the agents and $\mathbf{m} := \rho\mathbf{v}$ the flux. A potential mean-field game typically has the following formulation:

$$\min_{(\rho,\mathbf{m}) \in \mathcal{C}(\mu_0)} \mathcal{L}(\rho, \mathbf{m}) := \int_0^1 \int_\Omega \rho(\mathbf{x}, t) L\left(\mathbf{x}, \frac{\mathbf{m}(\mathbf{x}, t)}{\rho(\mathbf{x}, t)}\right) \mathrm{d}\mathbf{x}\mathrm{d}t + \int_0^1 \mathcal{F}_I(\rho(\cdot, t))\, \mathrm{d}t + \mathcal{F}_T(\rho(\cdot, 1)) \tag{3}$$

with the constraint set being

$$\mathcal{C}(\mu_0) := \{(\rho, \mathbf{m}) : \partial_t \rho + \nabla \cdot \mathbf{m} = 0, \rho(\cdot, 0) = \mu_0, \mathbf{m} \cdot \mathbf{n} = 0 \text{ for } \mathbf{x} \in \partial\Omega, \rho(\cdot, \cdot) \geqslant 0\}. \tag{4}$$

where $\mathbf{m}$ is the normal direction on the boundary $\partial\Omega$. It is clear to see that any pair of $(\rho, \mathbf{m}) \in \mathcal{C}(\mu_0)$ satisfies mass conservation and zero boundary flux condition with the initial density of ρ being μ_0. In this objective function, $L : \Omega \times \mathbb{R}^d \to \mathbb{R}$ models the dynamic cost, $\mathcal{F}_I : \mathcal{P}(\Omega) \to \mathbb{R}$ the interaction cost and $\mathcal{F}_T : \mathcal{P}(\Omega) \to \mathbb{R}$ the terminal cost.

To derive the optimality condition of (3), we introduce the Lagrangian multiplier ϕ and formulate the Lagrangian

$$\begin{aligned}\mathcal{A}(\rho,\mathbf{m},\phi): &= \mathcal{L}(\rho,\mathbf{m}) - \int_0^1\int_\Omega \phi(\mathbf{x},t)\left(\partial_t\rho(\mathbf{x},t)+\nabla\cdot\mathbf{m}(\mathbf{x},t)\right)\mathrm{d}\mathbf{x}\mathrm{d}t\\ &= \mathcal{L}(\rho,\mathbf{m}) + \int_0^1\int_\Omega \rho(\mathbf{x},t)\,\partial_t\phi(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\mathrm{d}t + \int_0^1\int_\Omega \mathbf{m}(\mathbf{x},t)\cdot\nabla\phi(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\mathrm{d}t\\ &\quad - \int_\Omega \phi(\mathbf{x},1)\,\rho(\mathbf{x},1)\,\mathrm{d}\mathbf{x} + \int_\Omega \phi(\mathbf{x},0)\,\mu_0(\mathbf{x})\,\mathrm{d}\mathbf{x},\end{aligned} \tag{5}$$

where the second equality is due to integration by part. The optimal solution solves the saddle point problem

$$\min_{\rho\geqslant 0,\mathbf{m}}\max_{\phi}\mathcal{A}(\rho,\mathbf{m},\phi). \tag{6}$$

When $L(\mathbf{x},\mathbf{v})$ is convex in $\mathbf{v}$, let the Legendre transformation of L be

$$H:\Omega\times\mathbb{R}^d\to\mathbb{R},(\mathbf{x},\mathbf{p})\mapsto\sup_{\mathbf{v}}\left\{-\langle\mathbf{p},\mathbf{v}\rangle - L(\mathbf{x},\mathbf{v})\right\}. \tag{7}$$

Then if $\rho>0$, the optimality condition of (3) is

$$\begin{cases} -\partial_t\phi(\mathbf{x},t)+H(\mathbf{x},\nabla\phi(\mathbf{x},t)) = \dfrac{\delta\mathcal{F}_I(\rho)}{\delta\rho}(\mathbf{x}), \quad \phi(\mathbf{x},1) = \dfrac{\delta\mathcal{F}_T(\rho)}{\delta\rho}(\mathbf{x}),\\ \partial_t\rho(\mathbf{x},t) - \nabla\cdot\left(\rho(\mathbf{x},t)\,\partial_{\mathbf{p}}H(\mathbf{x},\nabla\phi(\mathbf{x},t))\right) = 0, \quad \rho(\cdot,0)=\mu_0.\end{cases} \tag{8}$$

We use this forward-backward PDE system to explore the properties of the inverse problem later.

In this paper, we focus on the following two problems.

Problem 2.1 (crowd motion with obstacle). A common example comes from crowd motion [30], whose formulation is

$$\begin{aligned}\min_{(\rho,\mathbf{m})\in\mathcal{C}(\mu_0)}\mathcal{L}(\rho,\mathbf{m};b) := &\int_0^1\int_\Omega \frac{\|\mathbf{m}(\mathbf{x},t)\|_2^2}{2\rho(\mathbf{x},t)}\mathrm{d}\mathbf{x}\mathrm{d}t + \int_0^1\int_\Omega \rho(\mathbf{x},t)\,b(\mathbf{x})\,\mathrm{d}\mathbf{x}\mathrm{d}t\\ &+\gamma_I\int_0^1\int_\Omega \rho(\mathbf{x},t)\log\rho(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\mathrm{d}t\\ &+\gamma_T\int_\Omega \rho(\mathbf{x},t)\left(\log\rho(\mathbf{x},t)-\log\mu_1(\mathbf{x})\right)\mathrm{d}\mathbf{x}.\end{aligned} \tag{9}$$

Here the terminal cost is the KL divergence $\mathcal{F}_T(\rho(\cdot,1)) = \int_\Omega \rho(\mathbf{x},t)\left(\log\rho(\mathbf{x},t)-\log\mu_1(\mathbf{x})\right)\mathrm{d}\mathbf{x}$ which aims to match the terminal density $\rho(\cdot,1)$ to the desired density μ_1. The interaction cost contains two parts. The entropy term $\int_\Omega\rho(\mathbf{x},t)\log\rho(\mathbf{x},t)\mathrm{d}\mathbf{x}$ penalizes the aggregation of the density. And the obstacle term $\int_\Omega\rho(\mathbf{x},t)b(\mathbf{x})\mathrm{d}\mathbf{x}$ penalizes the mass going through the obstacle

$\mathbf{x}$ with larger value of $b(\mathbf{x})$. With the same initial density μ_0, different obstacle functions lead to different Nash Equilibrium. Assuming that we know everything in the objective function (9) except the obstacle function b, we aim to recover b from observations of the equilibrium $(\rho, \mathbf{m})$.

Problem 2.2 (non-Euclidean metric). It is also common to consider mean-field games on spaces with non-Euclidean metrics. If at each $\mathbf{x} \in \Omega, \Omega \subset \mathbb{R}^d$, there is a positive definite matrix $g(\mathbf{x}) \in S^d_{++}$ indicating the metric, then the mean-field game problem takes the form

$$\begin{aligned}\min_{(\rho,\mathbf{m})\in\mathcal{C}(\mu_0)} \mathcal{L}(\rho,\mathbf{m};g) := &\int_0^1 \int_\Omega \frac{\mathbf{m}(\mathbf{x},t)^\top g(\mathbf{x})\mathbf{m}(\mathbf{x},t)}{2\rho(\mathbf{x},t)}\mathrm{d}\mathbf{x}\mathrm{d}t \\ &+ \gamma_I \int_0^1 \int_\Omega \rho(\mathbf{x},t)\log\rho(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\mathrm{d}t \\ &+ \gamma_T \int_\Omega \rho(\mathbf{x},t)\left(\log\rho(\mathbf{x},t) - \log\mu_1(\mathbf{x})\right)\mathrm{d}\mathbf{x}.\end{aligned} \tag{10}$$

We also work on solving the metric g from the observations of the equilibrium $(\rho, \mathbf{m})$, assuming other terms in (10) are known.

In summary, we are interested in the mean-field game problem with the objective function

$$\begin{aligned}\mathcal{L}(\rho,\mathbf{m};g,b) := &\int_0^1 \int_\Omega \frac{\mathbf{m}(\mathbf{x},t)^\top g(\mathbf{x})\mathbf{m}(\mathbf{x},t)}{2\rho(\mathbf{x},t)}\mathrm{d}\mathbf{x}\mathrm{d}t + \int_0^1 \int_\Omega \rho(\mathbf{x},t)\,b(\mathbf{x})\,\mathrm{d}\mathbf{x}\mathrm{d}t \\ &+ \gamma_I \int_0^1 \int_\Omega \rho(\mathbf{x},t)\log\rho(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\mathrm{d}t + \gamma_T \int_\Omega \rho(\mathbf{x},1)\left(\log\rho(\mathbf{x},1) - \log\mu_1(\mathbf{x})\right)\mathrm{d}\mathbf{x}.\end{aligned} \tag{11}$$

We write $\mathcal{L}(\rho,\mathbf{m};g)$ when $b \equiv 0$ and $\mathcal{L}(\rho,\mathbf{m};b)$ when $g \equiv I_d$. With $\rho > 0$, the optimality condition for the problem

$$\min_{(\rho,\mathbf{m})\in\mathcal{C}(\mu_0)} \mathcal{L}(\rho,\mathbf{m};g,b), \tag{12}$$

is

$$\begin{cases} -\partial_t\phi(\mathbf{x},t) + \dfrac{1}{2}\left(\nabla\phi(\mathbf{x},t)\right)^\top \left(g(\mathbf{x})\right)^{-1}\nabla\phi(\mathbf{x},t) = \gamma_I\left(\log\rho(\mathbf{x},t)+1\right) + b(\mathbf{x}), \\ \qquad \phi(\mathbf{x},1) = \gamma_T\left(\log\rho(\mathbf{x},t) - \log\mu_1(\mathbf{x}) + 1\right), \\ \partial_t\rho(\mathbf{x},t) - \nabla\cdot\left(\rho(\mathbf{x},t)\left(g(\mathbf{x})\right)^{-1}\nabla\phi(\mathbf{x},t)\right) = 0, \quad \rho(\cdot,0) = \mu_0. \end{cases} \tag{13}$$

We call the potential mean-field games (3), as well as (9) and (10), the forward problem. In this paper, we aim to learn the unknown variables b, g from one or a set of observations of the Nash Equilibrium $\left\{\left(\widetilde{\rho}^n, \widetilde{\mathbf{m}}^n\right)\right\}_{n=1}^N$ that solve the forward problems, and we name this the inverse problem. Note that the forward problem has a convex objective function and linear constraint, while the optimality condition is nonlinear and nonconvex. To preserve the nice convex-linear structure of the forward problem, we formulate the inverse mean-field game as a bilevel optimization problem and treat the forward problem as the constraint.

Remark 2.3. While this paper mainly works on the inverse problem of problems 2.1 and 2.2, we emphasis that the bilevel formulation introduced in section 3.1 and the alternating gradient algorithm in section 4 are applicable to a broad class of inverse mean-field games, provided the cost function can be parameterized, either by values on a grid or by a neural network. More importantly, our convergence analysis holds for a very general class of mean-field games whose forward objective exhibits convexity with respect to the density ρ and momentum $\mathbf{m}$.

3. A bilevel formulation of inverse mean-field games

In this section, we first review the general formulation of a bilevel optimization problem, then provide the bilevel formulation of inverse mean-field games, as well as two concrete inverse problems that we would like to solve in this work. After that, we discretize the model for numerical implementation.

3.1. Bilevel formulation

The general formulation of a bilevel optimization problem is

$$\begin{aligned} &\min_{\xi\in\Xi} \quad u(\xi) := \mathcal{U}(\eta^*(\xi);\xi) \\ &\text{where } \eta^*(\xi) = \operatorname*{argmin}_{\eta\in H} \mathcal{L}(\eta;\xi). \end{aligned} \tag{14}$$

Here we consider linear constraint set $H = \{\eta \mid A\eta = c\}$ and convex set Ξ, where $A \in \mathbb{R}^{d_c\times d_\eta}, c \in \mathbb{R}^{d_c}$. $d_c < d_\eta$. The optimization problem over $\mathcal{U}$ is referred to as the upper-level problem and that over $\mathcal{L}$ as the lower-level problem. We formulate our inverse problems as bilevel optimization problems, with the upper-level objective being a combination of fidelity $\mathcal{D}_\rho, \mathcal{D}_\mathbf{m}$ and regularity $\mathcal{R}$, and the lower-level problem being the forward problem.

$$\begin{aligned} \min_{L\in\mathcal{C}_L, \mathcal{F}_I\in\mathcal{C}_{\mathcal{F}_I}} \quad & \mathcal{U}((\rho,\mathbf{m}),(\widetilde{\rho},\widetilde{\mathbf{m}});L,\mathcal{F}_I) := (\mathcal{D}_\rho(\rho,\widetilde{\rho}) + \mathcal{D}_\mathbf{m}(\mathbf{m},\widetilde{\mathbf{m}})) + \mathcal{R}(L,\mathcal{F}_I) \\ \text{s.t. } & (\rho,\mathbf{m}) := \operatorname*{argmin}_{(\rho,\mathbf{m})\in\mathcal{C}(\mu_0)} \mathcal{L}(\rho,\mathbf{m};L,\mathcal{F}_I). \end{aligned}$$

The dynamic cost L and interaction cost functional $\mathcal{F}_I$ are the upper-level variables and the density-flux pair $(\rho,\mathbf{m})$ is the lower-level variable. For convenience, we choose $\mathcal{D}_\rho(\rho,\widetilde{\rho}) = \frac{1}{2}\int_0^1\int_\Omega(\rho(\mathbf{x},t)-\widetilde{\rho}(\mathbf{x},t))^2\mathrm{d}\mathbf{x}\mathrm{d}t$ and $\mathcal{D}_\rho(\mathbf{m},\widetilde{\mathbf{m}}) = \frac{1}{2}\int_0^1\int_\Omega\|\mathbf{m}(\mathbf{x},t)-\widetilde{\mathbf{m}}(\mathbf{x},t)\|_2^2\mathrm{d}\mathbf{x}\mathrm{d}t$

We formulate the inverse problems of problem 2.1 and 2.2 as follows.

Problem 3.1 (the inverse problem of crowd motion (problem 2.1)). Let the regularity be $\mathcal{R}(b) = 0$. The inverse problem of (11) is

$$\begin{aligned} \min_{b\in\mathcal{C}_b} \quad & \mathcal{D}_\rho(\rho,\widetilde{\rho}) + \mathcal{D}_\mathbf{m}(\mathbf{m},\widetilde{\mathbf{m}}) \\ \text{s.t. } & (\rho,\mathbf{m}) := \operatorname*{argmin}_{(\rho,\mathbf{m})\in\mathcal{C}(\mu_0)} \mathcal{L}(\rho,\mathbf{m};b). \end{aligned} \tag{15}$$

Here $(\widetilde{\rho},\widetilde{\mathbf{m}}) = \operatorname{argmin}_{(\rho,\mathbf{m})\in\mathcal{C}(\mu_0)}\mathcal{L}(\rho,\mathbf{m};\widetilde{b})$ are the observed data with ground truth $\widetilde{b}$. Notice that for any constant $c\in\mathbb{R}$, if $(\widetilde{\rho},\widetilde{\mathbf{m}})$ minimizes $\mathcal{L}(\rho,\mathbf{m};\widetilde{b})$, then $(\widetilde{\rho},\widetilde{\mathbf{m}})$ also minimizes

$\mathcal{L}(\rho,\mathbf{m};\widetilde{b}+c)$. To remove the ambiguity, we restrict our focus to obstacle functions with zero integral, i.e.

$$\mathcal{C}_b := \left\{ b : \int_\Omega b(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0 \right\}. \tag{16}$$

Ideally, we expect $\mathrm{proj}_{\mathcal{C}_b}(\widetilde{b})$ to be the unique minimizer of the bilevel problem (15). We prove this unique identifiability property for the discretization of (15) in section 5.

Problem 3.2 (the inverse problem of unknown metric (problem 2.2)). Similarly, we have the bilevel formulation to recover the metric $\widetilde{g}$ from the data $(\widetilde{\rho},\widetilde{\mathbf{m}}) = \underset{(\rho,\mathbf{m})\in\mathcal{C}(\mu_0)}{\mathrm{argmin}} \mathcal{L}(\rho,\mathbf{m};\widetilde{g})$.

$$\begin{aligned} \min_{g\in\mathcal{C}_g} \quad & \mathcal{D}_\rho(\rho,\widetilde{\rho}) + \mathcal{D}_{\mathbf{m}}(\mathbf{m},\widetilde{\mathbf{m}}) + \mathcal{R}(g) \\ \text{s.t. } & (\rho,\mathbf{m}) := \underset{(\rho,\mathbf{m})\in\mathcal{C}(\mu_0)}{\mathrm{argmin}} \mathcal{L}(\rho,\mathbf{m};g). \end{aligned} \tag{17}$$

To make sure that g induces a metric on Ω, we set the constraint of g as

$$\mathcal{C}_g := \left\{ g : \Omega \to S^d_{++} : g(\mathbf{x}) \text{ are positive definite matrices}, \forall \mathbf{x}\in\Omega \right\}. \tag{18}$$

For one observation, if the density is zero in an open set, it means almost no players pass the region and it is impossible to obtain the exact information in that region. However multiple observations may complement the missing information, and therefore it is meaningful to consider the following inverse MFG with multiple observations.

Problem 3.3 (the inverse problem of unknown metric (problem 2.2) with multiple observations). Suppose that we have multiple observations of the Nash Equilibrium with a given $\widetilde{g}$ from different initial densities $\mu_0^n, n = 1,\dots,N$, i.e. $(\widetilde{\rho}^n,\widetilde{\mathbf{m}}^n) = \underset{(\rho,\mathbf{m})\in\mathcal{C}(\mu_0^n)}{\mathrm{argmin}} \mathcal{L}(\rho,\mathbf{m};\widetilde{g})$ for $n = 1,2,\dots,N$. Then we can solve the following bilevel optimization problem to recover the true metric

$$\begin{aligned} \min_{g\in\mathcal{C}_g} \quad & \sum_{n=1}^N \left(\mathcal{D}_\rho(\rho^n,\widetilde{\rho}^n) + \mathcal{D}_{\mathbf{m}}\left(\mathbf{m}^n,\widetilde{\mathbf{m}}^n\right) \right) + \mathcal{R}(g) \\ \text{s.t. } & \{(\rho^n,\mathbf{m}^n)\}_{n=1}^N := \underset{(\rho_n,\mathbf{m}_n)\in\mathcal{C}\left(\mu_0^n\right)}{\mathrm{argmin}} \sum_{n=1}^N \mathcal{L}(\rho_n,\mathbf{m}_n;g). \end{aligned} \tag{19}$$

The lower-level is equivalent to a concatenation of N forward problems since $(\rho^n,\mathbf{m}^n)$ are independent.

3.2. Discretization

We conduct numerical experiments on $\mathbb{R}^d$ with $d = 1,2$. Taking $d = 2$ as an example, we let $\Omega = [0,1]\times[0,1]$ and the space-time joint domain be $[0,1]^3$, and we write $\mathbf{m} = (m^x, m^y)$. We follow the discretization in [34], with which the discrete optimizer is consistent with the continuous optimizer under certain regularity conditions. To be precise, we equally divide $[0,1]$ into n_x, n_y, n_t parts, and each cube is of size $\Delta x\Delta y\Delta t$, with $\Delta x = \frac{1}{n_x}, \Delta y = \frac{1}{n_y}, \Delta t = \frac{1}{n_t}$. Let $x_i = (i-\frac{1}{2})\Delta x, y_i = (i-\frac{1}{2})\Delta y, t_i = (i-\frac{1}{2})\Delta t$, and $(f)_{i_x i_y i_t}$ approximates function f on the

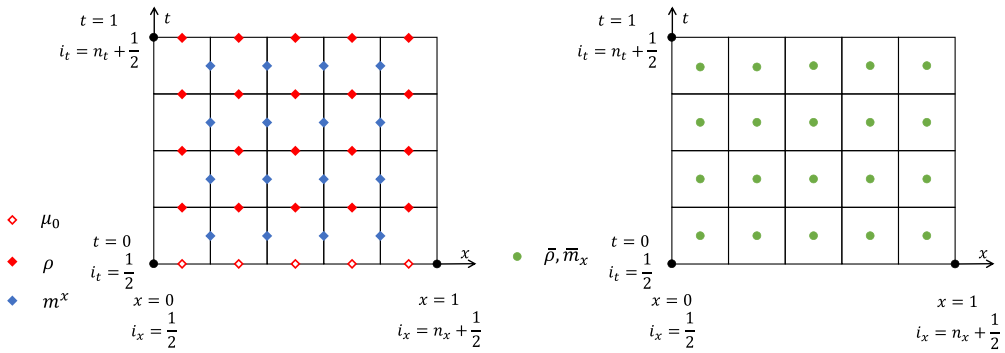

Figure 1. Illustrations of the staggered (left) and central (right) grids. Reproduced with permission from [33].

point $(x_{i_x}, y_{i_y}, t_{i_t})$. Similarly, $(f)_{i_x,i_y} \approx f(x_{i_x}, y_{i_y})$. We define $\mathcal{G}^{\rho}, \mathcal{G}^{m^x}$ and $\mathcal{G}^{m^y}$ as the sets of grid point indices on t-, x- and y-staggered grids, respectively, where

$$\mathcal{G}^{\rho} := \left\{ \left(i_x, i_y, i_t + \frac{1}{2} \right) : i_x = 1, \ldots, n_x, i_y = 1, \ldots, n_y, i_t = 1, \ldots, n_t \right\},$$
$$\mathcal{G}^{m^x} := \left\{ \left(i_x + \frac{1}{2}, i_y, i_t \right) : i_x = 1, \ldots, n_x - 1, i_y = 1, \ldots, n_y, i_t = 1, \ldots, n_t \right\},$$
$$\mathcal{G}^{m^y} := \left\{ \left(i_x, i_y + \frac{1}{2}, i_t \right) : i_x = 1, \ldots, n_x, i_y = 1, \ldots, n_y - 1, i_t = 1, \ldots, n_t \right\}.$$

Then we approximate the function ρ, m^x and m^y on t-, x- and y-staggered grids by $\rho_{\mathcal{G}^{\rho}}, m^x_{\mathcal{G}^{m^x}}$ and $m^y_{\mathcal{G}^{m^y}}$, respectively, i.e. $\rho_{\mathcal{G}^{\rho}} := \{(\rho)_{\mathbf{i}}\}_{\mathbf{i}\in\mathcal{G}^{\rho}} \in \mathbb{R}^{n_x n_y n_t}$, $m^x_{\mathcal{G}^{m^x}} = \{(m^x)_{\mathbf{i}}\}_{\mathbf{i}\in\mathcal{G}^{m^x}} \in \mathbb{R}^{(n_x-1)n_y n_t}$ and $m^y_{\mathcal{G}^{m^y}} = \{(m^y)_{\mathbf{i}}\}_{\mathbf{i}\in\mathcal{G}^{m^y}} \in \mathbb{R}^{n_x(n_y-1)n_t}$. We denote $\mathcal{G}^{\mathbf{m}} := \mathcal{G}^{m^x} \times \mathcal{G}^{m^y}$ as the concatenation of $\mathcal{G}^{m^x}, \mathcal{G}^{m^y}$ and $\mathbf{m}_{\mathcal{G}^{\mathbf{m}}} := \{m^x_{\mathcal{G}^{m^x}}, m^y_{\mathcal{G}^{m^y}}\}$ as the concatenation of $m^x_{\mathcal{G}^{m^x}}, m^y_{\mathcal{G}^{m^y}}$. We will omit the under scripts of grids wherever there is no ambiguity according to context. The left part of figure 1 illustrates the staggered grids and the corresponding ρ, m^x for $d = 1$.

We define the inner products on the staggered grids as

$$\langle \rho_1, \rho_2 \rangle_{\mathcal{G}^{\rho}} := \Delta x \Delta y \Delta t \sum_{\mathbf{i}\in\mathcal{G}^{\rho}} (\rho_1)_{\mathbf{i}} (\rho_2)_{\mathbf{i}},$$
$$\langle m^x_1, m^x_2 \rangle_{\mathcal{G}^{m^x}} := \Delta x \Delta y \Delta t \sum_{\mathbf{i}\in\mathcal{G}^{m^x}} (m^x_1)_{\mathbf{i}} (m^x_2)_{\mathbf{i}},$$
$$\langle m^y_1, m^y_2 \rangle_{\mathcal{G}^{m^y}} := \Delta x \Delta y \Delta t \sum_{\mathbf{i}\in\mathcal{G}^{m^y}} (m^y_1)_{\mathbf{i}} (m^y_2)_{\mathbf{i}},$$

and denote their induced norm as $\|\cdot\|_{\mathcal{G}^{\rho}}, \|\cdot\|_{\mathcal{G}^{m^x}}$ and $\|\cdot\|_{\mathcal{G}^{m^y}}$. Based on these, we approximate the discrepancy between lower-level minimizer and observed data $\mathcal{D}_{\rho}, \mathcal{D}_{\mathbf{m}}$ by the sum of element-wise differences on grids $\mathcal{D}_{\mathcal{G}^{\rho}}, \mathcal{D}_{\mathcal{G}^{\mathbf{m}}}$,

$$\begin{aligned} \mathcal{D}_{\mathcal{G}^{\rho}}(\rho, \widetilde{\rho}) &:= \frac{1}{2} \|\rho - \widetilde{\rho}\|^2_{\mathcal{G}^{\rho}} \\ \mathcal{D}_{\mathcal{G}^{\mathbf{m}}}(\mathbf{m}, \widetilde{\mathbf{m}}) &:= \frac{1}{2} \|m^x - \widetilde{m}^x\|^2_{\mathcal{G}^{m^x}} + \frac{1}{2} \|m^y - \widetilde{m}^y\|^2_{\mathcal{G}^{m^y}}. \end{aligned} \tag{20}$$

To compute the objective function, we consider the central grid (see the right plot in figure 1)

$$\mathcal{G}^{\phi} := \{(i_x, i_y, i_t) : i_x = 1, \ldots, n_x, i_y = 1, \ldots, n_y, i_t = 1, \ldots, n_t\}.$$

We define the inner product and induced norm on the central grid similarly and denote them as $\langle \cdot, \cdot \rangle_{\mathcal{G}^{\phi}}$ and $\| \cdot \|_{\mathcal{G}^{\phi}}$. With the interpolation operators, $\overline{\rho} = I_t(\rho; \mu_0)$, $\overline{m^x} = I_x(m^x), \overline{m^y} = I_y(m^y)$ meet on the central grid points:

$$(\overline{\rho})_{i_x i_y i_t} = (I_t(\rho; \mu_0))_{i_x i_y i_t} := \begin{cases} \frac{1}{2}\left((\mu_0)_{i_x, i_y} + (\rho)_{i_x, i_y, i_t + \frac{1}{2}}\right), & i_t = 1, \\ \frac{1}{2}\left((\rho)_{i_x, i_y, i_t - \frac{1}{2}} + (\rho)_{i_x, i_y, i_t + \frac{1}{2}}\right), & i_t = 2, \ldots, n_t. \end{cases}$$

$$(\overline{m^x})_{i_x i_y i_t} = (I_x(m^x))_{i_x i_y i_t} := \begin{cases} \frac{1}{2}(m^x)_{i_x + \frac{1}{2}, i_y, i_t}, & i_x = 1, \\ \frac{1}{2}\left((m^x)_{i_x - \frac{1}{2}, i_y, i_t} + (m^x)_{i_x + \frac{1}{2}, i_y, i_t}\right), & i_x = 2, \ldots, n_x - 1, \\ \frac{1}{2}(m^x)_{i_x - \frac{1}{2}, i_y, i_t}, & i_x = n_x. \end{cases}$$

$$(\overline{m^y})_{i_x i_y i_t} = (I_y(m^y))_{i_x i_y i_t} := \begin{cases} \frac{1}{2}(m^y)_{i_x, i_y + \frac{1}{2}, i_t}, & i_y = 1, \\ \frac{1}{2}\left((m^y)_{i_x, i_y - \frac{1}{2}, i_t} + (m^y)_{i_x, i_y + \frac{1}{2}, i_t}\right), & i_y = 2, \ldots, n_y - 1, \\ \frac{1}{2}(m^y)_{i_x, i_y - \frac{1}{2}, i_t}, & i_y = n_y. \end{cases}$$

Here, the definition of $\overline{m^x}$ on $i_x = 1, n_x$ and $\overline{m^y}$ on $i_y = 1, n_y$ are consistent with the zero-flux boundary condition in the continuous setting. The objective functions of the forward problem can therefore be approximated by

$$\begin{aligned} \mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; g, b) := \Delta x \Delta y \Delta t \sum_{\mathbf{i} \in \mathcal{G}^{\phi}} \left(\frac{(\overline{\mathbf{m}})_{\mathbf{i}}^{\top} (g)_{i_x, i_y} (\overline{\mathbf{m}})_{\mathbf{i}}}{2(\overline{\rho})_{\mathbf{i}}} + \gamma_I (\overline{\rho})_{\mathbf{i}} \log((\overline{\rho})_{\mathbf{i}}) \right) + \Delta x \Delta y \Delta t \sum_{\mathbf{i} \in \mathcal{G}^{\rho}} (\rho)_{\mathbf{i}} (b)_{i_x, i_y} \\ + \gamma_T \Delta x \Delta y \sum_{i_x = 1}^{n_x} \sum_{i_y = 1}^{n_y} (\rho)_{i_x, i_y, n_t + \frac{1}{2}} \left(\log(\rho)_{i_x, i_y, n_t + \frac{1}{2}} - \log(\mu_1)_{i_x, i_y} \right) \end{aligned} \tag{21}$$

where $\mathbf{m} = \{m^x, m^y\}$, $(\overline{\mathbf{m}})_{i_x i_y i_t}^{\top} := ((\overline{m^x})_{i_x i_y i_t}, (\overline{m^y})_{i_x i_y i_t})$ and the subscript of $\mathcal{L}_{\mathcal{G}}$ indicates the cost is defined on the discrete space. Similar to the continuous notation, we write $\mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; g)$ when $b = \mathbf{0}$ and $\mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; b)$ when $g \equiv 1$ $(d = 1)$ or $g \equiv I_2$ $(d = 2)$.

With this discretization, $\mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; g, b)$ preserves the following properties on $(\rho, \mathbf{m})$ from the continuous setting.

Lemma 3.4. *For* $\mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; g, b)$ *defined on* $(\rho_{\mathcal{G}^{\rho}}, m^x_{\mathcal{G}^{m^x}}, m^y_{\mathcal{G}^{m^y}}) \in \mathbb{R}^{n_x n_y n_t} \times \mathbb{R}^{(n_x - 1) n_y n_t} \times \mathbb{R}^{n_x (n_y - 1) n_t}$ *with* $\min_{\mathbf{i} \in \mathcal{G}^{\rho}} (\rho)_{\mathbf{i}} > 0$*, the following statements hold:*

1. *If* $\gamma_I, \gamma_T \geqslant 0$ *and* g_{i_x, i_y} *is positive definite for all* i_x, i_y*, then* $\mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; g, b)$ *is convex in* $\rho, \mathbf{m}$*.*
2. *In addition to* 1*, if we restrict the domain to* ρ *with* $\min_{\mathbf{i} \in \mathcal{G}^{\rho}} (\rho)_{\mathbf{i}} \geqslant \underline{c}_{\rho} > 0$*,* m^x *with* $\max_{\mathbf{i} \in \mathcal{G}^{m^x}} (|m^x_{\mathbf{i}}|) \leqslant \overline{c}_m$*, and* m^y *with* $\max_{\mathbf{i} \in \mathcal{G}^{m^y}} (|m^y_{\mathbf{i}}|) \leqslant \overline{c}_m$*, then* $\mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; g, b)$ *is Lipschitz smooth in* $\rho, \mathbf{m}$*.*
3. *In addition to* 1,2*, if we further restrict the domain to* $\rho \in \mathbb{R}^{n_x n_y n_t}$ *with* $\max_{\mathbf{i} \in \mathcal{G}^{\rho}} (\rho)_{\mathbf{i}} \leqslant \overline{c}_{\rho}$*, then for any* $\gamma_I, \gamma_T > 0$*,* $\mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; g, b)$ *is strongly convex in* $\rho, \mathbf{m}$*.*

We postpone the proof of the lemma in section 5 for better readability. The Lipschitz smoothness and strong convexity of the lower-level objective are important to guarantee the convergence of our alternating gradient algorithm, as detailed later in section 4.

Following the nature of the staggered grid, we choose a central difference scheme to approximate the differential operators

$$\left(D_t\left(\rho;\mu_0\right)\right)_{i_x i_y i_t} := \begin{cases} \frac{1}{\Delta t}\left(\left(\rho\right)_{i_x,i_y,i_t+\frac{1}{2}} - \left(\mu_0\right)_{i_x,i_y}\right), & i_t = 1, \\ \frac{1}{\Delta t}\left(\left(\rho\right)_{i_x,i_y,i_t+\frac{1}{2}} - \left(\rho\right)_{i_x,i_y,i_t-\frac{1}{2}}\right), & i_t = 2,\ldots,n_t. \end{cases}$$

$$\left(D_x\left(m^x\right)\right)_{i_x i_y i_t} := \begin{cases} \frac{1}{\Delta x}\left(m^x\right)_{i_x+\frac{1}{2},i_y,i_t}, & i_x = 1, \\ \frac{1}{\Delta x}\left(\left(m^x\right)_{i_x+\frac{1}{2},i_y,i_t} - \left(m^x\right)_{i_x-\frac{1}{2},i_y,i_t}\right), & i_x = 2,\ldots,n_x-1, \\ -\frac{1}{\Delta x}\left(m^x\right)_{i_x-\frac{1}{2},i_y,i_t}, & i_x = n_x. \end{cases}$$

$$\left(D_y\left(m^y\right)\right)_{i_x i_y i_t} := \begin{cases} \frac{1}{\Delta y}\left(m^y\right)_{i_x,i_y+\frac{1}{2},i_t}, & i_y = 1, \\ \frac{1}{\Delta y}\left(\left(m^y\right)_{i_x,i_y+\frac{1}{2},i_t} - \left(m^y\right)_{i_x,i_y-\frac{1}{2},i_t}\right), & i_y = 2,\ldots,n_y-1, \\ -\frac{1}{\Delta y}\left(m^y\right)_{i_x,i_y-\frac{1}{2},i_t}, & i_y = n_y. \end{cases}$$

Again, the definitions of D_x, D_y on $i_x = 1, n_x, i_y = 1, n_y$, respectively, are consistent with the zero-flux boundary condition. The discrete constraint set is

$$\mathcal{C}_{\mathcal{G}}\left(\mu_0\right) := \left\{\left(\rho,\mathbf{m}\right) : D_t\left(\rho;\mu_0\right) + D_x\left(m^x\right) + D_y\left(m^y\right) = \mathbf{0}\right\}. \tag{22}$$

Based on the above notations, we restate the inverse problems 3.1 and 3.2 in the discretized space. We intentionally write down the problems for more general cases with multiple pairs of training data as they will reduce to the case with a single pair of data by choosing $N = 1$.

Problem 3.5 (the discretization of the inverse crowd motion problem 3.1). The discretization of (15) has the following formulation

$$\begin{aligned} \min_{b\in\mathcal{C}_{\mathcal{G},b}} \quad & \sum_{n=1}^{N}\left(\mathcal{D}_{\mathcal{G}^\rho}\left(\rho^n,\widetilde{\rho}^n\right) + \mathcal{D}_{\mathcal{G}^{\mathbf{m}}}\left(\mathbf{m}^n,\widetilde{\mathbf{m}}^n\right)\right) \\ \text{s.t. } & \left(\rho^n,\mathbf{m}^n\right) := \operatorname*{argmin}_{(\rho,\mathbf{m})\in\mathcal{C}_{\mathcal{G}}\left(\mu_0^n\right)} \mathcal{L}_{\mathcal{G}}\left(\rho,\mathbf{m};b\right), n = 1,2,\ldots,N, \end{aligned} \tag{23}$$

where $\left(\widetilde{\rho}^n,\widetilde{\mathbf{m}}^n\right) = \operatorname{argmin}_{(\rho,\mathbf{m})\in\mathcal{C}_{\mathcal{G}}\left(\mu_0^n\right)} \mathcal{L}_{\mathcal{G}}(\rho,\mathbf{m};\widetilde{b})$ are the observed data and

$$\mathcal{C}_{\mathcal{G},b} := \left\{ b : \sum_{i_x,i_y}^{n_x,n_y} \left(b\right)_{i_x,i_y} = 0 \right\}. \tag{24}$$

Problem 3.6 (the discretization of the inverse metric problem 3.2). Similarly, given the data $\left(\widetilde{\rho}^n,\widetilde{\mathbf{m}}^n\right) = \operatorname*{argmin}_{(\rho,\mathbf{m})\in\mathcal{C}\left(\mu_0^n\right)} \mathcal{L}(\rho,\mathbf{m};\widetilde{g})$, we implement algorithms to solve

$$\begin{aligned} \min_{g\in\mathcal{C}_{\mathcal{G},g}} \quad & \sum_{n=1}^{N}\left(\mathcal{D}_{\mathcal{G}^\rho}\left(\rho^n,\widetilde{\rho}^n\right) + \mathcal{D}_{\mathcal{G}^{\mathbf{m}}}\left(\mathbf{m}^n,\widetilde{\mathbf{m}}^n\right)\right) + \mathcal{R}_{\mathcal{G}}\left(g\right) \\ \text{s.t. } & \left(\rho^n,\mathbf{m}^n\right) := \operatorname*{argmin}_{(\rho,\mathbf{m})\in\mathcal{C}_{\mathcal{G}}\left(\mu_0^n\right)} \mathcal{L}_{\mathcal{G}}\left(\rho,\mathbf{m};g\right), n = 1,2,\ldots,N, \end{aligned} \tag{25}$$

with the constraint of g being

$$\mathcal{C}_{\mathcal{G},g} := \left\{ g : (g)_{i_x,i_y} \in \mathbb{R}^{d\times d} \text{ are positive definite matrices}, i_x = 1,\dots,n_x, i_y = 1,\dots,n_y \right\}. \tag{26}$$

3.3. *Regularity and unique identifiability of the inverse problems*

At the end of this section, we state the regularity of the inverse problems 3.5 and 3.6 and the unique identifiability of the inverse crowd motion problem 3.5.

The regularity and unique identifiability of the inverse problem rely on the KKT system of the discretized forward problem

$$\min_{(\rho,\mathbf{m})\in\mathcal{C}(\mu_0)} \mathcal{L}_{\mathcal{G}}(\rho,\mathbf{m};g,b). \tag{27}$$

To write the KKT system in a concise way, we define the adjoint operators of I_x, I_y, I_t for any $\phi = \phi_{\mathcal{G}^\phi}$ on the central grid as

$$\begin{aligned}
\left(I_t^*(\phi)\right)_{i_x,i_y,i_t+\frac{1}{2}} &:= \begin{cases} \frac{1}{2}\left((\phi)_{i_x i_y i_t} + (\phi)_{i_x,i_y,i_t+1}\right), & i_t = 1,\dots,n_t-1, \\ \frac{1}{2}(\phi)_{i_x i_y i_t}, & i_t = n_t \end{cases} \\
\left(I_x^*(\phi)\right)_{i_x+\frac{1}{2},i_y,i_t} &:= \frac{1}{2}\left((\phi)_{i_x i_y i_t} + (\phi)_{i_x+1,i_y,i_t}\right), i_x = 1,\dots,n_x-1 \\
\left(I_y^*(\phi)\right)_{i_x,i_y+\frac{1}{2},i_t} &:= \frac{1}{2}\left((\phi)_{i_x i_y i_t} + (\phi)_{i_x,i_y+1,i_t}\right), i_y = 1,\dots,n_y-1.
\end{aligned}$$

And the adjoint operators of D_x, D_y, D_t as

$$\begin{aligned}
\left(D_t^*(\phi)\right)_{i_x,i_y,i_t+\frac{1}{2}} &:= \begin{cases} -\frac{1}{\Delta t}\left((\phi)_{i_x,i_y,i_t+1} - (\phi)_{i_x i_y i_t}\right), & i_t = 1,\dots,n_t-1 \\ \frac{1}{\Delta t}(\phi)_{i_x i_y i_t}, & i_t = n_t \end{cases} \\
\left(D_x^*(\phi)\right)_{i_x+\frac{1}{2},i_y,i_t} &:= \frac{1}{\Delta x}\left((\phi)_{i_x+1,i_y,i_t} - (\phi)_{i_x i_y i_t}\right), i_x = 1,\dots,n_x-1 \\
\left(D_y^*(\phi)\right)_{i_x,i_y+\frac{1}{2},i_t} &:= \frac{1}{\Delta y}\left((m^y)_{i_x,i_y+1,i_t} - (m^y)_{i_x i_y i_t}\right), i_y = 1,\dots,n_y-1.
\end{aligned}$$

The adjoint relation in the discretized space holds based on the definitions. To be precise, for the interpolation operators, we have

$$\begin{aligned}
\langle I_t(\rho;\mu_0),\phi\rangle_{\mathcal{G}^\phi} &= \langle \rho, I_t^*(\phi)\rangle_{\mathcal{G}^\rho} + \frac{1}{2}\sum_{i_x=1}^{n_x}\sum_{i_y=1}^{n_y}(\mu_0)_{i_x,i_y}(\phi)_{i_x,i_y,1} \\
\langle I_x(m^x),\phi\rangle_{\mathcal{G}^\phi} &= \langle m^x, I_x^*(\phi),\rangle_{\mathcal{G}^{m^x}} \\
\langle I_y(m^y),\phi\rangle_{\mathcal{G}^{m^y}} &= \langle m^y, I_y^*(\phi),\rangle_{\mathcal{G}^{m^y}}.
\end{aligned}$$

And for differential operators, we have

$$\begin{aligned}
\langle D_t(\rho;\mu_0),\phi\rangle_{\mathcal{G}^\phi} &= \langle \rho, D_t^*(\phi)\rangle_{\mathcal{G}^\rho} - \frac{1}{\Delta t}\sum_{i_x=1}^{n_x}\sum_{i_y=1}^{n_y}(\mu_0)_{i_x,i_y}(\phi)_{i_x,i_y,1} \\
\langle D_x(m^x),\phi\rangle_{\mathcal{G}^\phi} &= \langle m^x, D_x^*(\phi)\rangle_{\mathcal{G}^{m^x}} \\
\langle D_y(m^y),\phi\rangle_{\mathcal{G}^\phi} &= \langle m^y, D_y^*(\phi)\rangle_{\mathcal{G}^{m^y}}.
\end{aligned}$$

With the adjoint operators, we define the $\mathcal{Y}$ operators as following to describe the optimality condition for the forward problem,

$$\begin{cases} \mathbf{i} \in \mathcal{G}^{\rho}, i_t = 1, \ldots, n_t - 1, \\ (\mathcal{Y}_{\rho}(\rho, \mathbf{m}, \phi; g, b))_{\mathbf{i}} := -(D_t^*(\phi))_{\mathbf{i}} + \left(I_t^* \left(-\frac{(\overline{\mathbf{m}})^{\top} g \overline{\mathbf{m}}}{2\overline{\rho}^2} + \gamma_I (\log(\overline{\rho}) + 1) \right) \right)_{\mathbf{i}} + b_{i_x, i_y}, \\ \mathbf{i} \in \mathcal{G}^{\rho}, = n_t, \\ (\mathcal{Y}_{\rho}(\rho, \mathbf{m}, \phi; g, b))_{\mathbf{i}} := -(D_t^*(\phi))_{\mathbf{i}} + \left(I_t^* \left(-\frac{(\overline{\mathbf{m}})^{\top} g \overline{\mathbf{m}}}{2\overline{\rho}^2} + \gamma_I (\log(\overline{\rho}) + 1) \right) \right)_{\mathbf{i}} + b_{i_x, i_y} \\ \qquad + \frac{\gamma_T}{\Delta t} \left(\log(\rho_{\mathbf{i}}) - \log\left((\mu_1)_{i_x, i_y} \right) + 1 \right), \\ \mathbf{i} \in \mathcal{G}^{m^x}, \quad (\mathcal{Y}_{m^x}(\rho, \mathbf{m}, \phi; g, b))_{\mathbf{i}} := -(D_x^*(\phi))_{\mathbf{i}} + \left(I_x^* \left(\frac{g_{xx}\overline{m^x} + g_{xy}\overline{m^y}}{\overline{\rho}} \right) \right)_{\mathbf{i}}, \\ \mathbf{i} \in \mathcal{G}^{m^y}, \quad (\mathcal{Y}_{m^y}(\rho, \mathbf{m}, \phi; g, b))_{\mathbf{i}} := -\left(D_y^*(\phi)\right)_{\mathbf{i}} + \left(I_y^* \left(\frac{g_{xy}\overline{m^x} + g_{yy}\overline{m^y}}{\overline{\rho}} \right) \right)_{\mathbf{i}}, \\ \mathbf{i} \in \mathcal{G}^{\phi}, \quad (\mathcal{Y}_{\phi}(\rho, \mathbf{m}, \phi; g, b))_{\mathbf{i}} := (D_t(\rho; \mu_0) + D_x(m^x) + D_y(m^y))_{\mathbf{i}}. \end{cases} \tag{28}$$

$\mathcal{Y}_{\rho}, \mathcal{Y}_m^x, \mathcal{Y}_m^y, \mathcal{Y}_{\phi}$ are obtained by taking gradients on the Lagrangian of the forward problem (27). By viewing $\rho, \mathbf{m}, \phi, b$ and $\mathcal{Y}_{\rho}, \mathcal{Y}_m^x, \mathcal{Y}_m^y, \mathcal{Y}_{\phi}$ as long vectors and denoting $\mathcal{Y} := (\mathcal{Y}_{\rho}, \mathcal{Y}_{m^x}, \mathcal{Y}_{m^y}, \mathcal{Y}_{\phi})^{\top}$, we define a function $\mathcal{Y} : \mathbb{R}^{d_l} \times \mathbb{R}^{d_u} \to \mathbb{R}^{d_l}$ with $d_u = (\frac{d(d+1)}{2} + 1) n_x n_y$ corresponding to the dimension of (g, b) and $d_l = n_x n_y n_t + (n_x - 1) n_y n_t + n_x (n_y - 1) n_t + n_x n_y n_t$ to the dimension of ρ, m^x, m^y, ϕ. Since the constraint is linear, the optimizer of (27) satisfies the KKT condition. The formal statement is the following.

Lemma 3.7. *If $(\widetilde{\rho}, \widetilde{\mathbf{m}}) \in \mathcal{C}(\mu_0)$ is a minimizer of $\mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; \widetilde{g}, \widetilde{b})$, and $\min_{\mathbf{i} \in \mathcal{G}^{\rho}} \{\widetilde{\rho}_{\mathbf{i}}\} > 0$, then there exists $\widetilde{\phi} \in \mathbb{R}^{n_x n_y n_t}$ such that*

$$\mathcal{Y}\left(\widetilde{\rho}, \widetilde{\mathbf{m}}, \widetilde{\phi}; \widetilde{g}, \widetilde{b}\right) = \mathbf{0}. \tag{29}$$

With the discrete PDE description of the Nash Equilibrium, we state the regularity result for inverse problems 3.5 and 3.6.

Theorem 3.8 (regularity). *Assume that $(\widetilde{\rho}, \widetilde{\mathbf{m}})$ is the Nash Equilibrium given the metric $\widetilde{g}$, obstacle function $\widetilde{b}$ and $\gamma_{\mathrm{I}} > 0, \gamma_T > 0$, i.e. (27) holds, and that $\min_{\mathbf{i} \in \mathcal{G}^{\rho}} \widetilde{\rho}_{\mathbf{i}} > 0$, then there exists $r_u > 0$ and a radius r_u open ball $B_{r_u}(\widetilde{g}, \widetilde{b})$ centered at $(\widetilde{g}, \widetilde{b})$, and a mapping $\mathcal{T}$ defined on $B_{r_u}(\widetilde{g}, \widetilde{b})$ satisfying the following*

- *For any $(g, b) \in B_{r_u}(\widetilde{g}, \widetilde{b})$, there exist a unique $(\rho, \mathbf{m}, \phi) = \mathcal{T}(g, b) \in B_{r_l}(\widetilde{\rho}, \widetilde{\mathbf{m}}, \widetilde{\phi})$, a radius r_l open ball centered at $(\widetilde{\rho}, \widetilde{\mathbf{m}}, \widetilde{\phi})$, such that $(\rho, \mathbf{m}, \phi)$ solves the forward problem with $\mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; g, b)$.*
- *$\mathcal{T}(\widetilde{g}, \widetilde{b}) = (\widetilde{\rho}, \widetilde{\mathbf{m}}, \widetilde{\phi})$, $\mathcal{T}$ is of class C^1 and*

$$D\mathcal{T}(g, b) = -\left(D_{\rho, \mathbf{m}, \phi} \mathcal{Y}(\mathcal{T}(g, b); g, b))\right)^{-1} \left(D_b \mathcal{Y}(\mathcal{T}(g, b); g, b)\right), \text{ for all } (g, b) \in B_{r_u}(\widetilde{g}, \widetilde{b}). \tag{30}$$

In addition, we have the unique identifiability of the inverse crowd motion problem because the lower-level objective has a simple dependence on the obstacle b. To be concrete, by solving the inverse crowd motion problem 3.5, we uniquely recover the ground truth obstacle $\widetilde{b}$ up to a constant from only one good observation of the Nash Equilibrium.

Theorem 3.9 (unique identifiability). *Assume that $(\widetilde{\rho},\widetilde{\mathbf{m}})$ is the Nash Equilibrium given the obstacle function $\widetilde{b}$, i.e.*

$$(\widetilde{\rho},\widetilde{\mathbf{m}}) := \underset{(\rho,\mathbf{m})\in\mathcal{C}_{\mathcal{G}}(\mu_0)}{\operatorname{argmin}} \mathcal{L}_{\mathcal{G}}\left(\rho,\mathbf{m};\widetilde{b}\right), \tag{31}$$

and that $\min_{\mathbf{i}\in\mathcal{G}^{\rho}} \widetilde{\rho}_{\mathbf{i}} > 0$, then any minimizer b of the bilevel minimization problem

$$\begin{aligned} \min_{b} \quad & \mathcal{D}_{\mathcal{G}^{\rho}}(\rho,\widetilde{\rho}) + \mathcal{D}_{\mathcal{G}^{\mathbf{m}}}(\mathbf{m},\widetilde{\mathbf{m}}) \\ s.t. \quad & (\rho,\mathbf{m}) := \underset{(\rho,\mathbf{m})\in\mathcal{C}_{\mathcal{G}}(\mu_0)}{\operatorname{argmin}} \mathcal{L}_{\mathcal{G}}(\rho,\mathbf{m};b), \end{aligned} \tag{32}$$

has the form $b = \widetilde{b} + c$ where $c\in\mathbb{R}$ is a constant. This implies that $\widetilde{b}$ is the unique minimizer of the bilevel minimization problem (32) up to a constant.

The proofs are postponed to section 5. We close this section with some remarks on the theorems.

Remark 3.10 (numerical stability). While the unique identifiability theorem 3.9 holds without the entropy term and the regularity theorem 3.8, we emphasize that the entropy term and regularity theorem are meaningful for studying the numerical stability of the inverse problem. In fact, the entropy term guarantees the strong convexity of the objective function and thus the uniqueness of the forward problem. And it is important for the regularity theorem 3.8 to hold. The regularity argument states the differentiability of the forward optimizer with respect to the metric g and the obstacle b and reveals the rate of change. According to theorem 3.9, if the smallest singular value of $D\mathcal{T}(g,b)$ is large, then a small perturbation to $(\widetilde{\rho},\widetilde{\mathbf{m}})$ can still give a reasonable approximation of the ground truth $\widetilde{g},\widetilde{b}$. It is also worth noting that when $\min_{\mathbf{i}}\widetilde{\rho}_{\mathbf{i}}$ is close to 0, the condition number of the Jacobian matrix $D_{\rho,\mathbf{m},\phi}\mathcal{Y}(\widetilde{\rho},\widetilde{\mathbf{m}},\widetilde{\phi};b)$ in (30) can be extremely large. Therefore the Jacobian matrix $D_b(\rho,\mathbf{m},\phi)$ is close to singular, and the observation error may cause a failure to recover the ground truth obstacle.

Remark 3.11 (unique identifiability in the function space). Theorem 3.9 establishes the unique identifiability of the obstacle $b_{\mathcal{G}}\in\mathbb{R}^{n_x n_y}$ in the discretized finite-dimensional space. To prove the parallel result for the obstacle function $b:\Omega\to\mathbb{R}$ in the infinite-dimension space, it is subtle to choose the function space for $b,\rho,\mathbf{m}$, and ϕ. The function space is expected to be large enough to guarantee the existence of the lower-level optimizers $\rho^*(b),\mathbf{m}^*(b)$ for different b, and to guarantee the existence of the bilevel problem optimizer b^*. Meanwhile, the functions in the space require enough regularity for $\rho^*(b),\mathbf{m}^*(b)$ to be differentiable with respect to b. This is out of the scope of this paper. We refer interested readers to [24, 25, 29] for efforts in studying the unique identifiability in the infinite-dimensional space, where infinitely many pairs of training data are required.

Remark 3.12 (unique identifiability of the unknown metric). To establish the local unique identifiability of the metric as a corollary of the stability theorem 3.8, we need $D_g\mathcal{Y}(\widetilde{\rho},\widetilde{\mathbf{m}},\widetilde{\phi};g)$ to have full rank. However, for 1D metric, the rank of $D_g\mathcal{Y}(\widetilde{\rho},\widetilde{\mathbf{m}},\widetilde{\phi};g)$ depends on the data $\widetilde{\rho},\widetilde{\mathbf{m}},\widetilde{\phi}$, which is different from $D_b\mathcal{Y}(\widetilde{\rho},\widetilde{\mathbf{m}},\widetilde{\phi};b)$ being a constant. Therefore, we may not

uniquely recover the metric from the data. Besides the degenerated rank, while uniquely identifying g requires the knowledge of $\widetilde{\phi}$, we do not have $\widetilde{\phi}$ in our problem setting and this can also cause non-uniqueness of the inverse problem. By experiments in [8], the lack of information on $\widetilde{\phi}$ can be overcome by giving partial true information on the metric and incorporating regularity terms in the upper-level objective. For 2D metric, if we view g_{xx}, g_{xy}, g_{yy} as independent variables, then $D_g\mathcal{Y}(\widetilde{\rho}, \widetilde{\mathbf{m}}, \widetilde{\phi}; g)$ is not a full-rank matrix and theoretically there is no hope to uniquely recover the ground truth metric. If the metric $g_{\mathbf{i}} \in S^2_{++}$ has intrinsic structures such that the number of variables to determine the metric is $n_x n_y$ instead of $3n_x n_y$, numerically we recover the ground truth with a low error as shown by the numerical experiment in section 6.5. The numerical experiment in section 6.2.2 also shows that another way to resolve the ambiguity is to have multiple observations for more complete information in the region.

4. Alternating gradient method

In this section, we present the alternating gradient method (AGM) to solve the general bilevel optimization problem (14), as well as two inverse mean-field game problems 3.5 and 3.6.

4.1. Preliminary on AGM for bilevel optimization

The idea of the AGM is iteratively conducting gradient descent on the lower-level variable and the upper-level variable. To illustrate our algorithm, we first consider the following unconstrained bilevel problem

$$\begin{aligned} &\min_{\xi \in \mathbb{R}^{d_u}} \quad u(\xi) := \mathcal{U}(\eta^*(\xi); \xi) \\ &\text{where } \eta^*(\xi) = \operatorname*{argmin}_{\eta \in \mathbb{R}^{d_l}} \mathcal{L}(\eta; \xi). \end{aligned} \tag{33}$$

The computation of the lower-level gradient is straightforward. To obtain the upper-level gradient, we assume that $\mathcal{U}, \mathcal{L}$ are differentiable and denote the gradient operator with respect to their first and second entries as ∇_η, ∇_ξ. If for any given ξ, there exists a unique $\eta^*(\xi)$ solving the lower-level optimization problem and the function mapping ξ to its corresponding minimizer $\eta^*(\xi)$ is differentiable, then by chain rule, we have

$$\nabla u(\xi) = \nabla_\xi \eta^*(\xi)^\top \nabla_\eta \mathcal{U}(\eta^*(\xi); \xi) + \nabla_\xi \mathcal{U}(\eta^*(\xi); \xi), \tag{34}$$

with $\nabla_\xi \eta^*(\xi) = (\partial_{\xi_1} \eta^*(\xi), \ldots, \partial_{\xi_{d_u}} \eta^*(\xi)) \in \mathbb{R}^{d_l \times d_u}$ being the Jacobian matrix of η^*. We clarify that here $\nabla_\xi \mathcal{U}(\eta^*(\xi); \xi)$ is the gradient of $\mathcal{U}$ with respect to its second entry evaluated at $(\eta^*(\xi); \xi)$ without considering the dependence of η^* on ξ. Therefore $\nabla_\eta \mathcal{U}(\eta^*(\xi); \xi)$ and $\nabla_\xi \mathcal{U}(\eta^*(\xi); \xi)$ in (34) are easy to compute.

When the exact lower-level solution $\eta^*(\xi)$ is unavailable, the upper-level gradient $\nabla u(\xi)$ is inaccessible. However, we can approximate $\eta^*(\xi)$ and therefore approximate $\nabla_\xi u(\xi)$. Specifically, for ξ^{k_u} at the k_u-th iteration, we run K_l-step gradient descent of lower-level with stepsize τ_l to approximate $\eta^*(\xi^{k_u})$, i.e.

$$\begin{cases} \eta^{k_u,1} = \eta^{k_u}; \\ \eta^{k_u,k_l+1} = \eta^{k_u,k_l} - \tau_l \nabla_\eta \mathcal{L}\left(\eta^{k_u,k_l}; \xi^{k_u}\right), k_l = 1, \ldots, K_l; \\ \eta^{k_u+1} = \eta^{k_u,K_l+1}. \end{cases} \tag{35}$$

It is easy to see that $\eta^{k_u,k_l+1} = \eta^{k_u,k_l+1}(\xi^{k_u})(k_l = 1,\ldots,K_l)$ and $\eta^{k_u+1} = \eta^{k_u+1}(\xi^{k_u})$ are functions of ξ^{k_u}. We drop the dependence for notation conciseness and estimate the upper-level gradient $\nabla u(\xi^{k_u})$ by

$$\widehat{\nabla} u\left(\xi^{k_u}\right) := \left(\nabla_{\xi^{k_u}} \eta^{k_u+1}\right)^{\top} \nabla_{\eta} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) + \nabla_{\xi} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right), \tag{36}$$

where the η^{k_u} is a lower-level estimator of the lower-level optimizer $\eta^*(\xi^{k_u})$, and $(\nabla_{\xi^{k_u}} \eta^{k_u+1})_{ij} = \partial_{\xi_j^{k_u}} \eta_i^{k_u+1}$ estimates $\nabla_{\xi}\eta^*(\xi^{k_u})$ by unrolling the lower-level iterates through the chain rule. With the estimator in (36), we then update the upper-level variable by gradient descent with stepsize τ_u, i.e.

$$\xi^{k_u+1} = \xi^{k_u} - \tau_u \widehat{\nabla} u\left(\xi^{k_u}\right). \tag{37}$$

We summarize the algorithm in algorithm 1

Algorithm 1. General AGM for unconstrained bilevel optimization problem (33).

Initialization: ξ^1, η^1, stepsizes $\{\tau_u, \tau_l\}$
for $k_u = 1,2,\ldots,K_u$ **do**
  Initialize lower-level update by $\eta^{k_u,1} = \eta^{k_u}$.
  for $k_l = 1,2\cdots,K_l$ **do**
    lower-level gradient descent

$$\eta^{k_u,k_l+1} = \eta^{k_u,k_l} - \tau_l \nabla_{\eta} \mathcal{L}\left(\eta^{k_u,k_l}; \xi^{k_u}\right). \tag{38}$$

  end for
  Let the lower-level estimator be $\eta^{k_u+1} = \eta^{k_u,K_l+1}$ and compute $\widehat{\nabla} u(\xi^{k_u})$ by (36).
  Conduct upper-level gradient descent

$$\xi^{k_u+1} = \xi^{k_u} - \tau_u \widehat{\nabla} u\left(\xi^{k_u}\right). \tag{39}$$

end for

Remark 4.1 (error of unrolled differentiation). Equation (34) gives the exact value of the upper-level gradient. To obtain the unknown $\nabla_{\xi}\eta^*(\xi)$ in (34), we refer to the first-order optimality condition from the lower-level problem $\nabla_{\eta}\mathcal{L}(\eta^*(\xi);\xi) = \mathbf{0}$. We view $\nabla_{\eta}\mathcal{L}(\eta^*(\xi);\xi)$ as a vector-valued function of ξ, and its Jacobian matrix gives

$$\nabla_{\xi}\eta^*(\xi)^{\top} \nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi);\xi) + \nabla_{\xi\eta}\mathcal{L}(\eta^*(\xi);\xi) = \mathbf{0}, \tag{40}$$

where $(\nabla_{\xi\eta}\mathcal{L})_{ij}(\eta,\xi) = \partial_{\xi_i}\partial_{\eta_j}\mathcal{L}(\eta,\xi)$ and $(\nabla_{\eta\eta}\mathcal{L})_{ij}(\eta,\xi) = \partial_{\eta_i}\partial_{\eta_j}\mathcal{L}(\eta,\xi)$ are blocks of the Hessian matrix of $\mathcal{L}$. Therefore

$$\nabla_{\xi}\eta^*(\xi)^{\top} = -\nabla_{\xi\eta}\mathcal{L}(\eta^*(\xi);\xi)\left(\nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi);\xi)\right)^{-1}. \tag{41}$$

Plugging (41) into (34) gives the upper-level gradient

$$\nabla u(\xi) = \widehat{\nabla}_{\xi}\mathcal{U}(\eta^*(\xi);\xi), \tag{42}$$

where

$$\widehat{\nabla}_{\xi}\mathcal{U}(\eta;\xi) = \nabla_{\xi}\mathcal{U}(\eta;\xi) - \nabla_{\xi\eta}\mathcal{L}(\eta;\xi)\left(\nabla_{\eta\eta}\mathcal{L}(\eta;\xi)\right)^{-1}\nabla_{\eta}\mathcal{U}(\eta;\xi). \tag{43}$$

The gradient estimator (36) approximates the true gradient by approximating η^* by η^{k_u+1} and approximating $\left(\nabla_{\eta\eta}\mathcal{L}(\eta;\xi)\right)^{-1}$ by unrolling differentiation. A key to the convergence of the AGM algorithm is to control the error of unrolling differentiation. For unconstrained problems, [9, 14] proved that under sufficient smoothness assumptions, the errors of the approximations decrease as k_l increases. In lemma 5.3 of this paper, we study and prove the error can also be bounded for linear equality constrained lower-level problems.

4.2. AGM for inverse mean-field games

Building upon algorithm 1 for unconstrained bilevel optimization problems (33), we propose algorithm 2 to solve the constrained bilevel optimization problem (14) and its special cases in inverse mean-field game problems 3.1 and 3.2.

Algorithm 2. General AGM for (14).

Initialization: ξ^1, η^1, stepsizes $\{\tau_u, \tau_l\}$
for $k_u = 1, 2, \ldots, K_u$ **do**
  Initialize lower-level update by $\eta^{k_u,1} = \eta^{k_u}$.
  for $k_l = 1, 2\cdots, K_l$ **do**
    lower-level gradient descent

$$\eta^{k_u,k_l+1} = \operatorname*{proj}_{H}\left(\eta^{k_u,k_l} - \tau_l\nabla_{\eta}\mathcal{L}\left(\eta^{k_u,k_l};\xi^{k_u}\right)\right). \tag{44}$$

  end for
  Let the lower-level estimator be $\eta^{k_u+1} = \eta^{k_u,K_l+1}$ and compute $\widehat{\nabla}u(\xi^{k_u})$ by (36).
  Conduct upper-level projected gradient descent

$$\xi^{k_u+1} = \operatorname*{proj}_{\Xi}\left(\xi^{k_u} - \tau_u\widehat{\nabla}u\left(\xi^{k_u}\right)\right) \tag{45}$$

end for

Algorithm 2 applies the projected gradient descent to estimate the lower-level optimizer and to update the upper-level optimizer at each iteration. Precisely, by denoting the matrix form of the constraint (22) as $A\eta = c$, the projection to $H = \{\eta \mid A\eta = c\}$ is

$$\operatorname*{proj}_{H}(\eta) = \left(I - A^{\dagger}A\right)\eta + \eta_0,$$

where $A^{\dagger}$ is the Moore–Penrose inverse and η_0 is a fixed solution to $A\eta = c$. The projection operator is invariant to the lower-level objective and the number of iterations. As discussed

in [34], the main cost of the lower-level projected gradient descent is to compute the inverse of the discretized Laplacian operator $(AA^\top)^{-1}$, which can be solved efficiently using the fast cosine transform. We refer to section 3.2 in [34] for all detailed discussions. Since each step in projected gradient descent is explicit, it is possible to unroll the differentiation to estimate the upper-level gradient and thus conduct AGM for the constrained bilevel optimization problem. Although it is widely acknowledged in unconstrained bilevel optimization [9, 14] that the error arising from unrolling differentiation is controllable, rigorously adapting this approach to incorporate lower-level linear constraints is, to the best of our knowledge, unexplored. Lemma 5.3 in this paper investigates the error of this approximation, indicating that the gradient estimation error can be effectively bounded by the accuracy of the lower-level solution.

Remark 4.2 (the choice of lower-level (forward MFG) solver). The key of solving bilevel optimization problems with gradient-based method is to efficiently obtain the upper-level gradient estimator. Usually, this requires obtaining the lower-level optimizer $\eta^*(\xi)$ and the Jacobian matrix $\nabla_\xi \eta^*(\xi)$ through equation (34). While the lower-level optimizer $\eta^*(\xi)$ is easy to obtained from many forward MFG solvers, it is impractical to obtain the Jacobian matrix $\nabla_\xi \eta^*(\xi)$ because it is dense and of large size. Therefore, we implement our proximal gradient forward solver for K_l iterations to approximate $\eta^*(\xi)$ and use backpropagation to approximate $\nabla_\xi \eta^*(\xi)^\top \nabla_\eta \mathcal{U}(\eta^*(\xi);\xi)$. The proximal gradient solver for the lower-level problem [34] makes it easy and efficient to unroll the differentiation and estimate the upper-level gradient. It is worth emphasizing that this is not the case for other popular lower-level solvers, for example, primal-dual [27, 28], augmented Lagrangian [2, 3] and ADMM, because the implicit steps in ADMM and primal-dual methods make it impractical expensive and complicated to tracking the gradient.

The complexity of resolving the upper-level constraint is similar to a single-level optimization problem. In our cases, for the inverse crowd motion problem 3.5, the upper-level constraint set $\Xi = \mathcal{C}_{\mathcal{G},b}$ as defined in (24) is the set of matrices of size $n_x \times n_y$ with entry sum zero. And the projection is simply $\mathrm{proj}_{\mathcal{C}_{\mathcal{G},b}}(b) = \tilde{b}$, where $(\tilde{b})_{i_x,i_y} = (b)_{i_x,i_y} - \frac{1}{n_x n_y}\sum_{i_x,i_y}^{n_x,n_y}(b)_{i_x,i_y}$. And for the inverse metric problem 3.6, $\Xi = \mathcal{C}_{\mathcal{G},g}$, where $\mathcal{C}_{\mathcal{G},g}$ is defined in (26). We compute the projection $\tilde{g} := \mathrm{proj}_{\mathcal{C}_{\mathcal{G},g}}(g)$ pointwisely. To be specific, for $(g)_{i_x,i_y}$, we first compute its eigenvalue decomposition $(g)_{i_x,i_y} = Q\Lambda Q^{-1}$ where $\Lambda = \mathrm{diag}(\lambda_1,\lambda_2)$ and let $(\tilde{g})_{i_x,i_y} := Q\tilde{\Lambda}Q^{-1}$ where $\tilde{\Lambda} = \mathrm{diag}(\max(\lambda_1,\epsilon),\max(\lambda_2,\epsilon))$ with a pre-selected small positive value ϵ.

Different from our bilevel formulation and AGM algorithm, [7, 8] treat the forward MFG PDE system as the constraint of their optimization problem and apply primal-dual algorithm [5] to solve it. However, the nonlinear and nonconvex constraint makes it challenging to prove the algorithm convergence. On the contrary, our bilevel formulation takes advantage of the convex-linear structure of the forward MFG and we establish the following convergence theorem of our algorithm 2.

If the upper-level and lower-level objective functions satisfy the following regularity assumptions,

Assumption 1. Assume that $\mathcal{U}, \nabla\mathcal{U}, \nabla\mathcal{L}, \nabla^2\mathcal{L}$ is Lipschitz continuous with $\ell_{u,0}, \ell_{u,1}, \ell_{l,1}, \ell_{l,2}$, respectively.

Assumption 2. For any fixed ξ, assume that $\mathcal{L}(\eta;\xi)$ is μ_l-strongly convex with respect to η.

Assumption 3. Ξ is a linear constraint set $\Xi = \{\xi \mid B\xi = e\}$, and H and Ξ are nonempty.

then we have the following theorem.

Theorem 4.3. *Under assumptions 1–3, let* $\tau_l \leqslant \frac{1}{2\ell_{l,1}}, K_l = \mathcal{O}(\log K_u)$ *and* $\tau_u = \mathcal{O}(1)$*, then the iterates of algorithm 2 satisfy*

$$\frac{1}{K_u}\sum_{k_u=1}^{K_u} \|\xi^{k_u} - \operatorname*{proj}_{\Xi}\left(\xi^{k_u} - \nabla u\left(\xi^{k_u}\right)\right)\|^2 = \mathcal{O}\left(\frac{1}{K_u}\right) \tag{46}$$

where $\mathcal{O}$ *omits the log dependency.*

Let us define ϵ stationary point as $\|\xi - \operatorname{proj}_{\Xi}(\xi - \nabla u(\xi))\|^2 \leqslant \epsilon$, then theorem 4.3 states that algorithm 2 achieves ϵ stationary point by $\mathcal{O}(\epsilon^{-1})$ iterations. This matches the iteration complexity of the single-level projected gradient descent method. We postpone the proof in section 5.

Lemma 3.4 states that when $\rho, \mathbf{m}$ are bounded, and when the entropy in the objective function is non-zero ($\lambda > 0$), then our inverse problems 3.5 and 3.6 satisfy assumptions 1 and 2. Moreover, since the upper-level constraint set of the inverse crowd motion problem 3.5 is linear, assumption 3 is satisfied and theorem 4.3 guarantees the algorithm convergence when solving problem 3.5. For the inverse metric problem 3.6 where the upper-level constraint set is a convex cone, the convergence of the algorithm can be established similarly. However, the convergence rate is possibly different. We leave the study of the convergence rate for general upper-level constraints set to future research.

At the end of this section, we discuss how to unroll differentiation in practice.

Remark 4.4 (unroll differentiation in practice). Recall that in our problem, the lower-level variable $\eta = (\rho_{\mathcal{G}^\rho}, \mathbf{m}_{\mathcal{G}^{\mathbf{m}}})$ and the upper-level variable $\xi = (g_{\mathcal{G}}, b_{\mathcal{G}})$ are of size $\mathcal{O}(d^2 n_t n_x n_y)$. To obtain the upper-level gradient estimator (36), the computation of $\nabla_{\xi}\mathcal{U}\left(\eta^{k_u+1};\xi^{k_u}\right)$ is straightforward. But it is not practicable to directly formulate $\nabla_{\xi^{k_u}}\eta^{k_u+1}$ since the size of the Jacobian matrix is $\mathcal{O}(dn_t n_x n_y) \times \mathcal{O}(d^2 n_t n_x n_y)$ and the sparsity structure of the Jacobian matrix is not straightforward. Denote the gradient descent mapping $M(\eta;\xi) := \eta - \tau_l \nabla_{\eta}\mathcal{L}(\eta;\xi)$. Then the Jacobian of M, $\nabla M = (\nabla_{\eta}M, \nabla_{\xi}M) = (I - \tau\nabla_{\eta\eta}\mathcal{L}, -\tau\nabla_{\eta\xi}\mathcal{L})$ is sparse because the number of non-zero entries of $\nabla_{\eta\eta}\mathcal{L}$ and $\nabla_{\eta\xi}\mathcal{L}$ is $\mathcal{O}(dn_t n_x n_y)$. In practice, we avoid formulating the matrix $\nabla_{\xi^{k_u}}\eta^{k_u+1}$ by chain rule and the sparsity structure of ∇M. Specifically, let $P := I - A^{\dagger}A$ be the projection matrix, $\nabla_{\eta^{k_u,k_l}}\mathcal{U}\left(\eta^{k_u+1};\xi^{k_u}\right)$ be the gradient of $\mathcal{U}\left(\eta^{k_u+1};\xi^{k_u}\right)$ with respect to η^{k_u,k_l}, and $\nabla_{\xi^{k_u}}\eta^{k_u,k_l}$ be the Jacobian of η^{k_u,k_l} with respect to ξ^{k_u}, then we have the following relation by back-propagation

$$\begin{cases} \nabla_{\eta^{k_u,k_l+1}}\mathcal{U}\left(\eta^{k_u+1};\xi^{k_u}\right) = \nabla_{\eta}\mathcal{U}\left(\eta^{k_u+1};\xi^{k_u}\right), \\ \nabla_{\eta^{k_u,k_l}}\mathcal{U}\left(\eta^{k_u+1};\xi^{k_u}\right) = \left(\nabla_{\eta}M\left(\eta^{k_u,k_l};\xi^{k_u}\right)\right)^{\top} P \nabla_{\eta^{k_u,k_l+1}}\mathcal{U}\left(\eta^{k_u};\xi^{k_u}\right), \quad k_l = 1,\ldots,K_l. \end{cases} \tag{47}$$

Consequently, the upper-level gradient estimator is

$$
\begin{aligned}
\widehat{\nabla} u\left(\xi^{k_u}\right) &= \left(\nabla_{\xi^{k_u}} \eta^{k_u, K_l+1}\right)^{\top} \nabla_{\eta^{k_u, K_l+1}} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) + \nabla_{\xi} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) \\
&= \left(\nabla_{\eta} M\left(\eta^{k_u, K_l}; \xi^{k_u}\right) \nabla_{\xi^{k_u}} \eta^{k_u, K_l}\right)^{\top} P \nabla_{\eta^{k_u, K_l+1}} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) \\
&\quad + \left(\nabla_{\xi} M\left(\eta^{k_u, K_l}; \xi^{k_u}\right)\right)^{\top} P \nabla_{\eta^{k_u, K_l+1}} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) + \nabla_{\xi} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) \\
&\overset{\text{by (47)}}{=} \left(\nabla_{\xi^{k_u}} \eta^{k_u, K_l}\right)^{\top} \nabla_{\eta^{k_u, K_l}} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) \\
&\quad + \left(\nabla_{\xi} M\left(\eta^{k_u, K_l}; \xi^{k_u}\right)\right)^{\top} P \nabla_{\eta^{k_u, K_l+1}} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) + \nabla_{\xi} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) \\
&= \left(\nabla_{\eta} M\left(\eta^{k_u, K_l-1}; \xi^{k_u}\right) \nabla_{\xi^{k_u}} \eta^{k_u, K_l-1}\right)^{\top} P \nabla_{\eta^{k_u, K_l}} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) \\
&\quad + \left(\nabla_{\xi} M\left(\eta^{k_u, K_l-1}; \xi^{k_u}\right)\right)^{\top} P \nabla_{\eta^{k_u, K_l}} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) \\
&\quad + \left(\nabla_{\xi} M(\eta^{k_u, K_l}; \xi^{k_u})\right)^{\top} P \nabla_{\eta^{k_u, K_l+1}} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) + \nabla_{\xi} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) \\
&\overset{\text{by (47)}}{=} \left(\nabla_{\xi^{k_u}} \eta^{k_u, K_l-1}\right)^{\top} \nabla_{\eta^{k_u, K_l-1}} \left(\eta^{k_u+1}; \xi^{k_u}\right) \\
&\quad + \sum_{i=K_l-1}^{K_l} \left(\nabla_{\xi} M(\eta^{k_u, i}; \xi^{k_u})\right)^{\top} P \nabla_{\eta^{k_u, i+1}} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) + \nabla_{\xi} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) \\
&= \cdots \\
&\overset{(a)}{=} \sum_{i=1}^{K_l} \left(\nabla_{\xi} M(\eta^{k_u, i}; \xi^{k_u})\right)^{\top} P \nabla_{\eta^{k_u, i+1}} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right) + \nabla_{\xi} \mathcal{U}\left(\eta^{k_u+1}; \xi^{k_u}\right)
\end{aligned} \tag{48}
$$

where (a) is because that $\eta^{k_u,1}$ is independent of ξ^{k_u}. In this way, each term in the estimator can be computed by sparse matrix and vector multiplication.

5. Proofs of main theorems

In this section, we provide the proofs of main theorems. Theorem 3.8 shows that the observations of the Nash Equilibrium continuously depend on the unknown parameters. Theorem 3.9 states that with only one good observation of the Nash Equilibrium, we can uniquely recover the obstacle b up to a constant by solving the bilevel problem (23). This illustrates the effectiveness of our model. Lemma 3.4 and theorem 4.3 together guarantee that algorithm 2 converges to a stationary point to the bilevel problem (23) if the forward problem has enough regularity. This illustrates the effectiveness of our algorithm.

5.1. Proof of theorems 3.8 and 3.9

Recall that $\mathcal{Y}(\rho, \mathbf{m}, \phi; g, b) = \mathbf{0}$ gives the optimality condition. Denote the Jacobian matrix of $\mathcal{Y}$ as $\nabla \mathcal{Y} = ((\nabla_{\rho,\mathbf{m},\phi} \mathcal{Y})_{d_l \times d_l}, (\nabla_{g,b} \mathcal{Y})_{d_l \times d_u})$. The proof of the regularity theorem 3.8 is based on the implicit function theorem and the key is to show that the matrix $\nabla_{\rho,\mathbf{m},\phi} \mathcal{Y}$ is invertible at a good observation $(\widetilde{\rho}, \widetilde{\mathbf{m}}, \widetilde{\phi}; \widetilde{g}, \widetilde{b})$.

Lemma 5.1. *If* $\gamma_I > 0, \gamma_T > 0$ *and* $\min_{\mathbf{i}\in\mathcal{G}^\rho}\{\widetilde{\rho}_{\mathbf{i}}\} > 0$*, then* $\nabla_{\rho,\mathbf{m},\phi}\mathcal{Y}(\widetilde{\rho},\widetilde{\mathbf{m}},\widetilde{\phi};\widetilde{g},\widetilde{b})$ *is invertible.*

Proof. To prove the lemma is equivalent to showing that

$$\nabla_{\rho,\mathbf{m},\phi}\mathcal{Y}\left(\widetilde{\rho},\widetilde{\mathbf{m}},\widetilde{\phi};\widetilde{g},\widetilde{b}\right)(\delta_\rho,\delta_{\mathbf{m}},\delta_\phi) = \mathbf{0}, \tag{49}$$

if and only if $(\delta_\rho,\delta_{\mathbf{m}},\delta_\phi) = \mathbf{0}$. Here $\delta_{\mathbf{m}} := \{\delta_{m^x},\delta_{m^y}\}$. By definition,

$$\begin{aligned}\nabla_{\rho,\mathbf{m},\phi}\mathcal{Y}\left(\widetilde{\rho},\widetilde{\mathbf{m}},\widetilde{\phi};\widetilde{g},\widetilde{b}\right)(\delta_\rho,\delta_{\mathbf{m}},\delta_\phi) = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\Big(&\mathcal{Y}\left(\widetilde{\rho}+\epsilon\delta_\rho,\widetilde{\mathbf{m}}+\epsilon\delta_{\mathbf{m}},\widetilde{\phi}+\epsilon\delta_\phi;\widetilde{g},\widetilde{b}\right)\\ &-\mathcal{Y}\left(\widetilde{\rho},\widetilde{\mathbf{m}},\widetilde{\phi};\widetilde{g},\widetilde{b}\right)\Big).\end{aligned} \tag{50}$$

Therefore (49) is equivalent to

$$\begin{cases}
\mathbf{i}\in\mathcal{G}^\rho, i_t = 1,\dots,n_t-1,\\
-\left(D_t^*(\delta_\phi)\right)_{\mathbf{i}} + \left(I_t^*\left(-\dfrac{g_{xx}\overline{\widetilde{m}^x}+g_{xy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\overline{\delta_{m^x}} - \dfrac{g_{xy}\overline{\widetilde{m}^x}+g_{yy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\overline{\delta_{m^y}}\right.\right.\\
\qquad\left.\left. + \dfrac{\left(\overline{\widetilde{\mathbf{m}}}\right)^\top g\overline{\widetilde{\mathbf{m}}}}{\overline{\widetilde{\rho}}^3}\overline{\delta_\rho} + \dfrac{\gamma_I}{\overline{\widetilde{\rho}}}\overline{\delta_\rho}\right)\right)_{\mathbf{i}} = 0,\\
\mathbf{i}\in\mathcal{G}^\rho, = n_t,\\
-\left(D_t^*(\delta_\phi)\right)_{\mathbf{i}} + \left(I_t^*\left(-\dfrac{g_{xx}\overline{\widetilde{m}^x}+g_{xy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\overline{\delta_{m^x}} - \dfrac{g_{xy}\overline{\widetilde{m}^x}+g_{yy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\overline{\delta_{m^y}}\right.\right.\\
\qquad\left.\left. + \dfrac{\left(\overline{\widetilde{\mathbf{m}}}\right)^\top g\overline{\widetilde{\mathbf{m}}}}{\overline{\widetilde{\rho}}^3}\overline{\delta_\rho} + \dfrac{\gamma_I}{\overline{\widetilde{\rho}}}\overline{\delta_\rho}\right)\right)_{\mathbf{i}} + \dfrac{\gamma_T}{\Delta t(\widetilde{\rho})_{\mathbf{i}}}(\delta_\rho)_{\mathbf{i}} = 0,\\
\mathbf{i}\in\mathcal{G}^{m^x}, \quad -\left(D_x^*(\delta_\phi)\right)_{\mathbf{i}} + \left(I_x^*\left(\dfrac{g_{xx}}{\overline{\widetilde{\rho}}}\overline{\delta_{m^x}} + \dfrac{g_{xy}}{\overline{\widetilde{\rho}}}\overline{\delta_{m^y}} - \dfrac{g_{xx}\overline{\widetilde{m}^x}+g_{xy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\overline{\delta_\rho}\right)\right)_{\mathbf{i}} = 0,\\
\mathbf{i}\in\mathcal{G}^{m^y}, \quad -\left(D_y^*(\delta_\phi)\right)_{\mathbf{i}} + \left(I_y^*\left(\dfrac{g_{xy}}{\overline{\widetilde{\rho}}}\overline{\delta_{m^x}} + \dfrac{g_{yy}}{\overline{\widetilde{\rho}}}\overline{\delta_{m^y}} - \dfrac{g_{xy}\overline{\widetilde{m}^x}+g_{yy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\overline{\delta_\rho}\right)\right)_{\mathbf{i}} = 0,\\
\mathbf{i}\in\mathcal{G}^\phi, \quad \left(D_t(\delta_\rho;\mathbf{0}) + D_x(\delta_{m^x}) + D_y(\delta_{m^y})\right)_{\mathbf{i}} = 0.
\end{cases} \tag{51}$$

Note that $\widetilde{\rho},\widetilde{\mathbf{m}},\widetilde{\phi}$ are viewed as constants with respect to $(\delta_\rho,\delta_{\mathbf{m}},\delta_\phi)$ in the system. It clear that the system 51 is linear in $(\delta_\rho,\delta_{\mathbf{m}},\delta_\phi)$ and therefore (49) holds if $(\delta_\rho,\delta_{\mathbf{m}},\delta_\phi) = \mathbf{0}$. If both

$(\delta_\rho, \delta_{\mathbf{m}}, \delta_\phi)$ and $(\delta'_\rho, \delta'_{\mathbf{m}}, \delta'_\phi)$ satisfy (49), then by plugging them into (51) and subtracting, we have

$$\begin{cases} \mathbf{i} \in \mathcal{G}^\rho, i_t = 1, \ldots, n_t - 1, \\ -\left(D_t^*\left(\delta_\phi - \delta'_\phi\right)\right)_{\mathbf{i}} + \left(I_t^* \left(-\frac{g_{xx}\overline{\widetilde{m}^x} + g_{xy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\left(\overline{\delta_{m^x}} - \overline{\delta'_{m^x}}\right) \right.\right. \\ \qquad\qquad - \frac{g_{xy}\overline{\widetilde{m}^x} + g_{yy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\left(\overline{\delta_{m^y}} - \overline{\delta'_{m^y}}\right) \\ \qquad\qquad \left.\left. + \frac{\left(\overline{\widetilde{\mathbf{m}}}\right)^\top g \overline{\widetilde{\mathbf{m}}}}{\overline{\widetilde{\rho}}^3}\left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right) + \frac{\gamma_I}{\overline{\widetilde{\rho}}}\left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right) \right) \right)_{\mathbf{i}} = 0, \\ \mathbf{i} \in \mathcal{G}^\rho, = n_t, \\ -\left(D_t^*\left(\delta_\phi - \delta'_\phi\right)\right)_{\mathbf{i}} + \left(I_t^* \left(-\frac{g_{xx}\overline{\widetilde{m}^x} + g_{xy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\left(\overline{\delta_{m^x}} - \overline{\delta'_{m^x}}\right) \right.\right. \\ \qquad\qquad - \frac{g_{xy}\overline{\widetilde{m}^x} + g_{yy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\left(\overline{\delta_{m^y}} - \overline{\delta'_{m^y}}\right) \\ \qquad\qquad \left.\left. + \frac{\left(\overline{\widetilde{\mathbf{m}}}\right)^\top g \overline{\widetilde{\mathbf{m}}}}{\overline{\widetilde{\rho}}^3}\left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right) + \frac{\gamma_I}{\overline{\widetilde{\rho}}}\left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right) \right) \right)_{\mathbf{i}} + \frac{\gamma_T}{\Delta t(\widetilde{\rho})_{\mathbf{i}}}\left(\delta_\rho - \delta'_\rho\right)_{\mathbf{i}} = 0, \end{cases} \tag{52}$$

$$\begin{cases} \mathbf{i} \in \mathcal{G}^{m^x}, \quad -\left(D_x^*\left(\delta_\phi - \delta'_\phi\right)\right)_{\mathbf{i}} + \left(I_x^* \left(\frac{g_{xx}}{\overline{\widetilde{\rho}}}\left(\overline{\delta_{m^x}} - \overline{\delta'_{m^x}}\right) + \frac{g_{xy}}{\overline{\widetilde{\rho}}}\left(\overline{\delta_{m^y}} - \overline{\delta'_{m^y}}\right) \right.\right. \\ \qquad\qquad \left.\left. - \frac{g_{xx}\overline{\widetilde{m}^x} + g_{xy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right) \right) \right)_{\mathbf{i}} = 0, \\ \mathbf{i} \in \mathcal{G}^{m^y}, \quad -\left(D_y^*\left(\delta_\phi - \delta'_\phi\right)\right)_{\mathbf{i}} + \left(I_y^* \left(\frac{g_{xy}}{\overline{\widetilde{\rho}}}\left(\overline{\delta_{m^x}} - \overline{\delta'_{m^x}}\right) + \frac{g_{yy}}{\overline{\widetilde{\rho}}}\left(\overline{\delta_{m^y}} - \overline{\delta'_{m^y}}\right) \right.\right. \\ \qquad\qquad \left.\left. - \frac{g_{xy}\overline{\widetilde{m}^x} + g_{yy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2}\left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right) \right) \right)_{\mathbf{i}} = 0, \end{cases} \tag{53}$$

and

$$\mathbf{i} \in \mathcal{G}^\phi, \quad \left(D_t\left(\delta_\rho - \delta'_\rho; \mathbf{0}\right) + D_x\left(\delta_{m^x} - \delta'_{m^x}\right) + D_y\left(\delta_{m^y} - \delta'_{m^y}\right)\right)_{\mathbf{i}} = 0. \tag{54}$$

Pointwisely multiplying (52) with $(\delta_\rho - \delta'_\rho)$ and summing over $\mathcal{G}^\rho$ gives us

$$\begin{aligned}
&-\left\langle \delta_\rho - \delta'_\rho, D_t^* \left(\delta_\phi - \delta'_\phi\right)\right\rangle_{\mathcal{G}^\rho} \\
&\quad - \left\langle \delta_\rho - \delta'_\rho, I_t^* \left(\frac{g_{xx}\overline{\widetilde{m}^x} + g_{xy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2} \left(\overline{\delta_{m^x}} - \overline{\delta'_{m^x}}\right)\right)\right\rangle_{\mathcal{G}^\rho} - \left\langle \delta_\rho - \delta'_\rho, I_t^* \left(\frac{g_{xy}\overline{\widetilde{m}^x} + g_{yy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2} \left(\overline{\delta_{m^y}} - \overline{\delta'_{m^y}}\right)\right)\right\rangle_{\mathcal{G}^\rho} \\
&\quad + \left\langle \delta_\rho - \delta'_\rho, I_t^* \left(\frac{\left(\overline{\widetilde{\mathbf{m}}}\right)^\top g \overline{\widetilde{\mathbf{m}}}}{\overline{\widetilde{\rho}}^3} \left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right)\right)\right\rangle_{\mathcal{G}^\rho} + \left\langle \delta_\rho - \delta'_\rho, I_t^* \left(\frac{\gamma_I}{\overline{\widetilde{\rho}}} \left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right)\right)\right\rangle_{\mathcal{G}^\rho} \\
&\quad + \Delta x \Delta y \sum_{n_x=1}^{n_x} \sum_{n_y=1}^{n_y} \frac{\gamma_T}{(\widetilde{\rho})_{i_x,i_y,n_t}} \left(\delta_\rho - \delta'_\rho\right)^2_{i_x,i_y,n_t} = 0.
\end{aligned} \tag{55}$$

Similarly (53) and (54) imply

$$\begin{aligned}
&-\left\langle \delta_{m^x} - \delta'_{m^x}, D_x^* \left(\delta_\phi - \delta'_\phi\right)\right\rangle_{\mathcal{G}^{m^x}} \\
&\quad + \left\langle \delta_{m^x} - \delta'_{m^x}, I_x^* \left(\frac{g_{xx}}{\overline{\widetilde{\rho}}} \left(\overline{\delta_{m^x}} - \overline{\delta'_{m^x}}\right) + \frac{g_{xy}}{\overline{\widetilde{\rho}}} \left(\overline{\delta_{m^y}} - \overline{\delta'_{m^y}}\right)\right)\right\rangle_{\mathcal{G}^{m^x}} \\
&\quad - \left\langle \delta_{m^x} - \delta'_{m^x}, I_x^* \left(\frac{g_{xx}\overline{\widetilde{m}^x} + g_{xy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2} \left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right)\right)\right\rangle_{\mathcal{G}^{m^x}} = 0,
\end{aligned} \tag{56}$$

and

$$\begin{aligned}
&-\left\langle \delta_{m^y} - \delta'_{m^y}, D_y^* \left(\delta_\phi - \delta'_\phi\right)\right\rangle_{\mathcal{G}^{m^y}} \\
&\quad + \left\langle \delta_{m^y} - \delta'_{m^y}, I_y^* \left(\frac{g_{xy}}{\overline{\widetilde{\rho}}} \left(\overline{\delta_{m^x}} - \overline{\delta'_{m^x}}\right) + \frac{g_{yy}}{\overline{\widetilde{\rho}}} \left(\overline{\delta_{m^y}} - \overline{\delta'_{m^y}}\right)\right)\right\rangle_{\mathcal{G}^{m^y}} \\
&\quad - \left\langle \delta_{m^y} - \delta'_{m^y}, I_y^* \left(\frac{g_{xy}\overline{\widetilde{m}^x} + g_{yy}\overline{\widetilde{m}^y}}{\overline{\widetilde{\rho}}^2} \left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right)\right)\right\rangle_{\mathcal{G}^{m^y}} = 0,
\end{aligned} \tag{57}$$

and

$$\left\langle \delta_\phi - \delta'_\phi, D_t\left(\delta_\rho - \delta'_\rho; \mathbf{0}\right)\right\rangle_{\mathcal{G}^\phi} + \left\langle \delta_\phi - \delta'_\phi, D_x\left(\delta_{m^x} - \delta'_{m^x}\right)\right\rangle_{\mathcal{G}^\phi} + \left\langle \delta_\phi - \delta'_\phi, D_y\left(\delta_{m^y} - \delta'_{m^y}\right)\right\rangle_{\mathcal{G}^\phi} = 0. \tag{58}$$

Next, we add (55)–(58) and combine terms with the same components in groups. The first group is

$$\begin{aligned}
&-\left\langle \delta_\rho - \delta'_\rho, D_t^* \left(\delta_\phi - \delta'_\phi\right)\right\rangle_{\mathcal{G}^\rho} - \left\langle \delta_{m^x} - \delta'_{m^x}, D_x^* \left(\delta_\phi - \delta'_\phi\right)\right\rangle_{\mathcal{G}^{m^x}} - \left\langle \delta_{m^y} - \delta'_{m^y}, D_y^* \left(\delta_\phi - \delta'_\phi\right)\right\rangle_{\mathcal{G}^{m^y}} \\
&\quad + \left\langle \delta_\phi - \delta'_\phi, D_t\left(\delta_\rho - \delta'_\rho; \mathbf{0}\right)\right\rangle_{\mathcal{G}^\phi} + \left\langle \delta_\phi - \delta'_\phi, D_x\left(\delta_{m^x} - \delta'_{m^x}\right)\right\rangle_{\mathcal{G}^\phi} + \left\langle \delta_\phi - \delta'_\phi, D_y\left(\delta_{m^y} - \delta'_{m^y}\right)\right\rangle_{\mathcal{G}^\phi},
\end{aligned} \tag{59}$$

and by the adjoint relation between D_t, D_t^*, this group sums to 0. The second group consists of

$$\left\langle \delta_\rho - \delta'_\rho, I_t^* \left(\frac{\gamma_I}{\overline{\widetilde{\rho}}} \left(\overline{\delta_\rho} - \overline{\delta'_\rho}\right)\right)\right\rangle_{\mathcal{G}^\rho} + \Delta x \Delta y \sum_{n_x=1}^{n_x} \sum_{n_y=1}^{n_y} \frac{\gamma_T}{(\widetilde{\rho})_{i_x,i_y,n_t}} \left(\delta_\rho - \delta'_\rho\right)^2_{i_x,i_y,n_t}. \tag{60}$$

And the sum is equal to $\gamma_I \| \frac{\overline{\delta_\rho} - \overline{\delta'_\rho}}{\overline{\widetilde{\rho}}^{1/2}} \|^2_{\mathcal{G}^\phi} + \Delta x \Delta y \sum_{n_x=1}^{n_x} \sum_{n_y=1}^{n_y} \frac{\gamma_T}{(\widetilde{\rho})_{i_x,i_y,n_t}} (\delta_\rho - \delta'_\rho)^2_{i_x,i_y,n_t}$. The rest terms form the last group and sum to

$$\sum_{\mathbf{i} \in \mathcal{G}^\phi} \frac{1}{\overline{\widetilde{\rho}}^3_{\mathbf{i}}} \left\| \left(\left(\overline{\delta_\rho} - \overline{\delta'_\rho} \right) \overline{\widetilde{\mathbf{m}}} - \overline{\widetilde{\rho}} \left(\overline{\delta_{\mathbf{m}}} - \overline{\delta'_{\mathbf{m}}} \right) \right)_{\mathbf{i}} \right\|^2_{g_{\mathbf{i}}}, \tag{61}$$

where $\|\mathbf{v}_{\mathbf{i}}\|^2_{g_{\mathbf{i}}} = (\mathbf{v})^\top_{\mathbf{i}} g_{\mathbf{i}} \mathbf{v}_{\mathbf{i}}$. Overall, adding (55)–(58) gives

$$\begin{aligned} &\gamma_I \left\| \frac{\overline{\delta_\rho} - \overline{\delta'_\rho}}{\overline{\widetilde{\rho}}^{1/2}} \right\|^2_{\mathcal{G}^\phi} + \Delta x \Delta y \sum_{n_x=1}^{n_x} \sum_{n_y=1}^{n_y} \frac{\gamma_T}{(\widetilde{\rho})_{i_x,i_y,n_t}} \left(\delta_\rho - \delta'_\rho \right)^2_{i_x,i_y,n_t} \\ &+ \sum_{\mathbf{i} \in \mathcal{G}^\phi} \frac{1}{\overline{\widetilde{\rho}}^3_{\mathbf{i}}} \left\| \left(\left(\overline{\delta_\rho} - \overline{\delta'_\rho} \right) \overline{\widetilde{\mathbf{m}}} - \overline{\widetilde{\rho}} \left(\overline{\delta_{\mathbf{m}}} - \overline{\delta'_{\mathbf{m}}} \right) \right)_{\mathbf{i}} \right\|^2_{g_{\mathbf{i}}} = 0. \end{aligned} \tag{62}$$

We conclude that each term in (62) is zero since they are non-negative and sum to zero. Combining $(\delta_\rho)_{i_x,i_y,n_t} = (\delta'_\rho)_{i_x,i_y,n_t}$ and $\overline{\delta_\rho} = \overline{\delta'_\rho}$ gives $\delta_\rho = \delta'_\rho$. Consequently, $\overline{\delta_{m^x}} = \overline{\delta'_{m^x}}$ and $\overline{\delta_{m^y}} = \overline{\delta'_{m^y}}$. Because I_x, I_y are full rank linear operators, $\delta_{m^x} = \delta'_{m^x}$ and $\delta_{m^y} = \delta'_{m^y}$. Based on $\delta_\rho = \delta'_\rho, \delta_{\mathbf{m}} = \delta'_{\mathbf{m}}$, (52) and (53) lead to $\delta_\phi = \delta'_\phi$. Therefore (49) has unique solution $(\rho, \mathbf{m}, \phi) = \mathbf{0}$, i.e. $\nabla_{\rho,\mathbf{m},\phi} \mathcal{Y}(\widetilde{\rho}, \widetilde{\mathbf{m}}, \widetilde{\phi}; \widetilde{b})$ is invertible. □

With lemma 5.1, we apply implicit function theorem to $\mathcal{Y}$ at $(\widetilde{\rho}, \widetilde{\mathbf{m}}, \widetilde{\phi}; \widetilde{g}, \widetilde{b})$ and then the regularity theorem 3.8 is true.

Next, we prove the unique identifiability theorem 3.9 for inverse obstacle problem 3.5.

Proof of theorem 3.9. Since the upper-level objective is non-negative and equals 0 when $b = \widetilde{b}$, any minimizer b of the bilevel minimization problem satisfies

$$(\widetilde{\rho}, \widetilde{\mathbf{m}}) = \underset{(\rho,\mathbf{m}) \in \mathcal{C}_{\mathcal{G}}(\mu_0)}{\operatorname{argmin}} \mathcal{L}_{\mathcal{G}}(\rho, \mathbf{m}; b), \tag{63}$$

and by lemma 3.7, there exists ϕ such that $\mathcal{Y}(\widetilde{\rho}, \widetilde{\mathbf{m}}, \phi; b) = \mathbf{0}$. Assume that b' is a minimizer, $b' \neq \widetilde{b}$, and

$$\mathcal{Y}\left(\widetilde{\rho}, \widetilde{\mathbf{m}}, \widetilde{\phi}; \widetilde{b}\right) = \mathcal{Y}(\widetilde{\rho}, \widetilde{\mathbf{m}}, \phi'; b') = \mathbf{0},$$

then

$$\mathcal{Y}\left(\widetilde{\rho}, \widetilde{\mathbf{m}}, \widetilde{\phi}; \widetilde{b}\right) - \mathcal{Y}(\widetilde{\rho}, \widetilde{\mathbf{m}}, \phi'; b') = \mathbf{0},$$

which is equivalent to

$$\begin{cases} \mathbf{i} \in \mathcal{G}^\rho, -\left(D_t^* \left(\phi' - \widetilde{\phi} \right) \right)_{\mathbf{i}} + \left((b')_{i_x,i_y} - \left(\widetilde{b} \right)_{i_x,i_y} \right) = 0, \\ \mathbf{i} \in \mathcal{G}^{m^x}, \left(D_x^* \left(\phi' - \widetilde{\phi} \right) \right)_{\mathbf{i}} = 0, \\ \mathbf{i} \in \mathcal{G}^{m^y}, \left(D_y^* \left(\phi' - \widetilde{\phi} \right) \right)_{\mathbf{i}} = 0. \end{cases} \tag{64}$$

The equation on $\mathcal{G}^\rho$ gives $(\phi' - \widetilde{\phi})_{i_x i_y i_t} = (n_t - i_t + 1)(b' - \widetilde{b})_{i_x,i_y}$. Plugging in equations on $\mathcal{G}^{m^x}, \mathcal{G}^{m^y}$, we have $(b' - \widetilde{b})_{i_x,i_y} = c$ where c is a constant for different i_x, i_y. □

5.2. *Proof of theorem 4.3 and lemma 3.4*

In this section, we provide the nonasymptotic analysis for AGM on general constrained bilevel optimization (14). We follow conventional notations in bilevel optimization by using commas to separate lower-level and upper-level variables, i.e. $\mathcal{L}(\eta,\xi)=\mathcal{L}(\eta;\xi),\mathcal{U}(\eta,\xi)=\mathcal{U}(\eta;\xi)$.

Recall that the lower level constraint is $H=\{\eta \mid A\eta=c\}$. Denote the singular value decomposition of A as $A=U\Sigma V^{\top}$, where

$$\Sigma=\begin{bmatrix}\Sigma_1 & 0\\ 0 & 0\end{bmatrix}\in\mathbb{R}^{d_c\times d_\eta},$$

$U=[U_1\ U_2], V=[V_1\ V_2], U\in\mathbb{R}^{d_c\times d_c}, V\in\mathbb{R}^{d_\eta\times d_\eta}$ are orthogonal matrix and $U_1\in\mathbb{R}^{d_c\times r}, V_1\in\mathbb{R}^{d_\eta\times r}$ are the submatrix corresponds to full rank diagonal submatrix $\Sigma_1\in\mathbb{R}^{r\times r}$. Then V_2 is the orthogonal basis of $\mathrm{Ker}(A):=\{\eta \mid A\eta=0\}$. Let $\eta_0\in H$ be a feasible lower-level solution, then the lower-level update is equivalent to

$$\eta^{k_u,1}=\eta^{k_u};\ \ \eta^{k_u,k_l+1}=V_2V_2^{\top}\left(\eta^{k_u,k_l}-\tau_l\nabla_\eta\mathcal{L}\left(\eta^{k_u,k_l},\xi^{k_u}\right)\right)+\eta_0;\ \ \eta^{k_u+1}=\eta^{k_u,K_l+1}. \tag{65}$$

With η^{k_u+1} approximating $\eta^*(\xi^{k_u})$, we approximate the lower-level gradient with

$$\widehat{\nabla}u\left(\xi^{k_u}\right):=\nabla_\xi\mathcal{U}\left(\eta^{k_u+1},\xi^{k_u}\right)+\left(\nabla_{\xi^{k_u}}\eta^{k_u+1}\right)^{\top}\nabla_\eta\mathcal{U}\left(\eta^{k_u+1},\xi^{k_u}\right) \tag{66}$$

and $\nabla_{\xi^{k_u}}\eta^{k_u+1}$ is obtained by unrolling the lower-level iterates

$$\begin{cases}\nabla_{\xi^{k_u}}\eta^{k_u,1}=\mathbf{0},\\ \nabla_{\xi^{k_u}}\eta^{k_u,k_l+1}=V_2V_2^{\top}\nabla_{\xi^{k_u}}\eta^{k_u,k_l}-\tau_lV_2V_2^{\top}\left(\nabla_{\eta\xi}\mathcal{L}\left(\eta^{k_u,k_l},\xi^{k_u}\right)+\nabla_{\eta\eta}\mathcal{L}\left(\eta^{k_u,k_l},\xi^{k_u}\right)\nabla_{\xi^{k_u}}\eta^{k_u,k_l}\right)\\ \qquad\qquad=V_2V_2^{\top}\left(I-\tau_l\nabla_{\eta\eta}\mathcal{L}\left(\eta^{k_u,k_l},\xi^{k_u}\right)\right)\nabla_{\xi^{k_u}}\eta^{k_u,k_l}-\tau_lV_2V_2^{\top}\nabla_{\eta\xi}\mathcal{L}\left(\eta^{k_u,k_l},\xi^{k_u}\right),k_l=1,\ldots,K_l.\end{cases} \tag{67}$$

To prove the convergence, we first present the regularity of the lower-level optimizer established in [32]. To be self-contained, we also provide its proof.

Lemma 5.2 (The regularity of lower-level optimizer). *Under assumptions 1 and 2, $\eta^*(\xi)$ is differentiable with respect to ξ with the following gradient*

$$\nabla\eta^*(\xi)=-V_2\left(V_2^{\top}\nabla_{\eta\eta}\mathcal{L}\left(\eta^*(\xi),\xi\right)V_2\right)^{-1}V_2^{\top}\nabla_{\eta\xi}\mathcal{L}\left(\eta^*(\xi),\xi\right)$$

where V_2 is the orthogonal basis of $\mathrm{Ker}(A)$. *Therefore, $\eta^*(\xi)$ is L_η-Lipschitz continuous and $L_{\eta\xi}$ smooth with*

$$L_\eta:=\frac{\ell_{l,1}}{\mu_l}=\mathcal{O}(\kappa),\qquad L_{\eta\xi}:=\frac{\ell_{l,2}\left(1+\frac{\ell_{l,1}}{\mu_l}\right)^2}{\mu_l}=\mathcal{O}\left(\kappa^3\right).$$

Proof. First, we prove the differentiability and compute the Jacobian matrix. We choose a fixed η_0 satisfying $A\eta_0=c$. Using the aforementioned SVD of A, the constraint set $H=\{\eta_0+V_2z \mid z\in\mathbb{R}^{d_\eta-r}\}$. Letting $\mathcal{L}_z(z,\xi):=\mathcal{L}(\eta_0+V_2z,\xi)$ and $z^*(\xi)=\arg\min_z\mathcal{L}_z(z,\xi)$, we have $\eta^*(\xi)=\eta_0+V_2z^*(\xi)$. By optimality condition, $z^*(\xi)$ satisfies

$$\nabla_z\mathcal{L}_z\left(z^*(\xi),\xi\right)=V_2^{\top}\nabla_\eta\mathcal{L}\left(\eta_0+V_2z^*(\xi),\xi\right)=0. \tag{68}$$

Since

$$\nabla_{zz}\mathcal{L}_z\left(z^*\left(\xi\right),\xi\right)=V_2^\top\nabla_{\eta\eta}\mathcal{L}\left(\eta_0+V_2z^*\left(\xi\right),\xi\right)V_2 \tag{69}$$

and by strong convexity of $\mathcal{L}$ with respect to η, $\nabla_{zz}\mathcal{L}_z(z^*(\xi),\xi)$ is invertible. By implicit function theorem, $z^*(\xi)$ is differentiable with respect to ξ. As a consequence, $\eta^*(\xi)$ is differentiable with respect to ξ. Taking the gradient with respect to ξ on both sides of (68) gives us

$$\begin{aligned}0&=\nabla_{\xi\eta}\mathcal{L}\left(\eta_0+V_2z^*\left(\xi\right),\xi\right)V_2+\left(\nabla_\xi z^*\left(\xi\right)^\top V_2^\top\right)\nabla_{\eta\eta}\mathcal{L}\left(\eta_0+V_2z^*\left(\xi\right),\xi\right)V_2\\&=\nabla_{\xi\eta}\mathcal{L}\left(\eta_0+V_2z^*\left(\xi\right),\xi\right)V_2+\nabla_\xi z^*\left(\xi\right)^\top V_2^\top\nabla_{\eta\eta}\mathcal{L}\left(\eta_0+V_2z^*\left(\xi\right),\xi\right)V_2.\end{aligned}$$

Then, we have (cf $\nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi),\xi)=\nabla_{\eta\eta}\mathcal{L}(\eta_0+V_2z^*(\xi),\xi)$)

$$\nabla z^*\left(\xi\right)=-\left(V_2^\top\nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi\right),\xi\right)V_2\right)^{-1}V_2^\top\nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi\right),\xi\right) \tag{70}$$

and as a result,

$$\begin{aligned}\nabla\eta^*\left(\xi\right)&=V_2\nabla z^*\left(\xi\right)\\&=-V_2\left(V_2^\top\nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi\right),\xi\right)V_2\right)^{-1}V_2^\top\nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi\right),\xi\right).\end{aligned}$$

Next, utilizing the fact that V_2 is the orthogonal matrix, we know $\mu_l I\preceq V_2^\top\nabla_{\eta\eta}\mathcal{L}(\eta,\xi)V_2$. Therefore, we have for any ξ,η,

$$V_2\left(V_2^\top\nabla_{\eta\eta}\mathcal{L}\left(\eta,\xi\right)V_2\right)^{-1}V_2^\top\preceq\frac{1}{\mu_l}I. \tag{71}$$

As a result, $\nabla\eta^*(\xi)$ is bounded by

$$\|\nabla\eta^*\left(\xi\right)\|\leqslant\|V_2\left(V_2^\top\nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi\right),\xi\right)V_2\right)^{-1}V_2^\top\|\|\nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi\right),\xi\right)\|\leqslant\frac{\ell_{l,1}}{\mu_l}=L_\eta$$

which implies $\eta^*(\xi)$ is L_η Lipschitz continuous.

Finally, we aim to prove the smoothness of $\eta^*(\xi)$. For any ξ_1 and ξ_2, we have

$$\begin{aligned}&\|\nabla\eta^*(\xi_1)-\nabla\eta^*(\xi_2)\|\\&\quad=\|V_2\left(V_2^\top\nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi_1),\xi_1)V_2\right)^{-1}V_2^\top\nabla_{\eta\xi}\mathcal{L}(\eta^*(\xi_1),\xi_1)\\&\qquad-V_2\left(V_2^\top\nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi_2),\xi_2)V_2\right)^{-1}V_2^\top\nabla_{\eta\xi}\mathcal{L}(\eta^*(\xi_2),\xi_2)\|\\&\quad\leqslant\|V_2B_1^{-1}V_2^\top\|\|\nabla_{\eta\xi}\mathcal{L}(\eta^*(\xi_1),\xi_1)-\nabla_{\eta\xi}\mathcal{L}(\eta^*(\xi_2),\xi_2))\|\\&\qquad+\|V_2(B_1^{-1}-B_2^{-1})V_2^\top\|\|\nabla_{\eta\xi}\mathcal{L}(\eta^*(\xi_2),\xi_2)\|\\&\quad\overset{(a)}{\leqslant}\frac{1}{\mu_l}\|\nabla_{\eta\xi}\mathcal{L}(\eta^*(\xi_1),\xi_1)-\nabla_{\eta\xi}\mathcal{L}(\eta^*(\xi_2),\xi_2)\|\\&\qquad+\frac{\ell_{l,1}}{\mu_l^2}\|\nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi_1),\xi_1)-\nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi_2),\xi_2)\|\\&\quad\overset{(b)}{\leqslant}\frac{\ell_{l,2}(1+\frac{\ell_{l,1}}{\mu_l})^2}{\mu_l}\|\xi_1-\xi_2\|\end{aligned} \tag{72}$$

where $B_1 = V_2^\top \nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi_1),\xi_1)V_2$ and $B_2 = V_2^\top \nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi_2),\xi_2)V_2$, (a) comes from (71) and the following fact:

$$\begin{aligned}
&V_2\left(B_1^{-1}-B_2^{-1}\right)V_2^\top \\
&\quad = V_2 B_1^{-1}\left(B_2-B_1\right)B_2^{-1}V_2^\top \\
&\quad = V_2 B_1^{-1}\left(\left(V_2^\top \nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi_2\right),\xi_2\right)V_2\right)-\left(V_2^\top \nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi_1\right),\xi_1\right)V_2\right)\right)B_2^{-1}V_2^\top \\
&\quad = V_2 B_1^{-1}V_2^\top\left(\nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi_2\right),\xi_2\right)-\nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi_1\right),\xi_1\right)\right)V_2 B_2^{-1}V_2^\top
\end{aligned}$$

and (b) comes from

$$\begin{aligned}
\|\nabla^2\mathcal{L}\left(\eta^*\left(\xi_1\right),\xi_1\right)-\nabla^2\mathcal{L}\left(\eta^*\left(\xi_2\right),\xi_2\right)\| &\leqslant \ell_{l,2}\left[\|\xi_1-\xi_2\|+\|\eta^*\left(\xi_1\right)-\eta^*\left(\xi_2\right)\|\right] \\
&\leqslant \ell_{l,2}\left(1+\frac{\ell_{l,1}}{\mu_l}\right)\|\xi_1-\xi_2\|.
\end{aligned}$$

□

In algorithm 2, we approximate $\nabla\eta^*(\xi)$ by unrolling the differentiation. The following lemma investigates the error of this approximation in constrained bilevel problems for the first time, indicating that the gradient estimation error can be effectively bounded by the accuracy of the lower-level solution.

Lemma 5.3 (Error of unrolling differentiation). *Suppose that assumptions 1–3 hold and choose* $\tau_l \leqslant \frac{1}{2\ell_{l,1}}$, *the error of implicit gradient estimator can be bounded by*

$$\|\nabla\eta^*\left(\xi^{k_u}\right)-\nabla_{\xi^{k_u}}\eta^{k_u+1}\|^2 \leqslant 2\left(1-\tau_l\mu_l\right)^{2K_l+2}+2C_{K_l}C_l^2\|\eta^*\left(\xi^{k_u}\right)-\eta^{k_u}\|^2$$

where $C_l^2 := (1+\frac{\ell_{l,1}}{\mu_l})\ell_{l,2}^2(\frac{2}{\mu_l^2}+\frac{3}{2\ell_{l,1}^2})$ *and* C_{K_l} *is the upper bound of* $K_l(1-\tau_l\mu_l)^{K_l-1}$ *and is finite.*

Proof. According to (67), we know that $\nabla_{\xi^{k_u}}\eta^{k_u,1}=0$ and

$$\nabla_{\xi^{k_u}}\eta^{k_u,k_l+1} = V_2V_2^\top\left(I-\tau_l\nabla_{\eta\eta}\mathcal{L}\left(\eta^{k_u,k_l},\xi^{k_u}\right)\right)\nabla_{\xi^{k_u}}\eta^{k_u,k_l}-\tau_l V_2V_2^\top\nabla_{\eta\xi}\mathcal{L}\left(\eta^{k_u,k_l},\xi^{k_u}\right).$$

For any given ξ^{k_u}, we can define an auxiliary sequence $\{w^{k_l}\}_{k_l=0}^{\infty}$ and $w^* := \lim_{K_l\to\infty} w^{K_l}$, where $w^1=0$ and

$$w^{k_l+1} = V_2V_2^\top\left(I-\tau_l\nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right)\right)w^{k_l}-\tau_l V_2V_2^\top\nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right). \tag{73}$$

We can see that (67) and (73) only differ in $\eta^*(\xi^{k_u})$ and η^{k_u,k_l}. For the sequence w^{k_l}, we can calculate the explicit form of w^{K_l+1} as

$$\begin{aligned}
w^{K_l+1} &= \sum_{s=0}^{K_l}\left(V_2V_2^\top-\tau_l V_2V_2^\top\nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right)\right)^s\left(-\tau_l V_2V_2^\top\nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right)\right) \\
&= \sum_{s=0}^{K_l}\left(V_2V_2^\top-\tau_l V_2V_2^\top\nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right)V_2V_2^\top\right)^s\left(-\tau_l\nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right)\right) \\
&= \sum_{s=0}^{K_l}\left(V_2\left(I-\tau_l V_2^\top\nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right)V_2\right)V_2^\top\right)^s\left(-\tau_l\nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=0}^{K_l} V_2 \left(I - \tau_l V_2^\top \nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right) V_2\right)^s V_2^\top \left(-\tau_l \nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right)\right) \\
&= V_2 \left(\sum_{s=0}^{K_l} \left(I - \tau_l V_2^\top \nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi^{k_u}),\xi^{k_u}) V_2\right)^s\right) V_2^\top \left(-\tau_l \nabla_{\eta\xi}\mathcal{L}(\eta^*(\xi^{k_u}),\xi^{k_u})\right) \qquad (74)
\end{aligned}
$$

where the first equality comes from unrolling (73), the second and the fourth equality are due to $(V_2V_2^\top)^s = V_2V_2^\top$. Let $D := I - \tau_l V_2^\top \nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi^{k_u}),\xi^{k_u})V_2$. When $\tau_l < \frac{2}{\ell_{l,1}}$, the operator norm of D satisfies $\|D\| < 1$, the limit $\sum_{s=0}^{+\infty} D^s := \lim_{K_l\to+\infty}\sum_{s=0}^{K_l} D^s = (I-D)^{-1} = (\tau_l V_2^\top \nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi^{k_u}),\xi^{k_u})V_2)^{-1}$. Therefore, the limit point of w^{k_l} is equal to $\nabla\eta^*(\xi^{k_u})$ since

$$
\begin{aligned}
w^* := \lim_{K_l\to\infty} w^{K_l} &= V_2 \left(\sum_{s=0}^{\infty} \left(I - \tau_l V_2^\top \nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right) V_2\right)^s\right) V_2^\top \left(-\tau_l \nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right)\right) \\
&= V_2 \left(\tau_l V_2^\top \nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right) V_2\right)^{-1} V_2^\top \left(-\tau_l \nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right)\right) \\
&= -V_2 \left(V_2^\top \nabla_{\eta\eta}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right) V_2\right)^{-1} V_2^\top \nabla_{\eta\xi}\mathcal{L}\left(\eta^*\left(\xi^{k_u}\right),\xi^{k_u}\right) = \nabla\eta^*\left(\xi^{k_u}\right). \qquad (75)
\end{aligned}
$$

Moreover, the error by finite-step approximation can be bounded by

$$
\left\|(I-D)^{-1} - \sum_{s=0}^{K_l} D^s\right\| = \left\|\sum_{s=K_l+1}^{\infty} D^s\right\| \leqslant \sum_{s=K_l+1}^{\infty} \|D\|^s = \frac{\|D\|^{K_l+1}}{1-\|D\|}.
$$

Since $(1-\tau_l\ell_{l,1})I \preceq D = I - \tau_l V_2^\top \nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi^{k_u}),\xi^{k_u})V_2 \preceq (1-\tau_l\mu_l)I$ and according to (75) and (74), we know that if $\tau_l \leqslant \frac{1}{2\ell_{l,1}}$,

$$
\|w^{K_l+1} - \nabla\eta^*\left(\xi^{k_u}\right)\| \leqslant \tau_l\ell_{l,1}\frac{(1-\tau_l\mu_l)^{K_l+1}}{1-\tau_l\ell_{l,1}} \leqslant (1-\tau_l\mu_l)^{K_l+1}. \qquad (76)
$$

Next, we aim to bound the distance between $\nabla_{\xi^{k_u}}\eta^{k_u,k_l}$ and the auxiliary sequence w^{k_l}. For any k_l, according to (67) and (73), we have

$$
\begin{aligned}
\|\nabla_{\xi^{k_u}}\eta^{k_u,k_l+1} - w^{k_l+1}\|^2 = &\| \left(V_2V_2^\top - \tau_l V_2V_2^\top \nabla_{\eta\eta}\mathcal{L}(\eta^{k_u,k_l},\xi^{k_u})\right)\nabla_{\xi^{k_u}}\eta^{k_u,k_l} - \tau_l V_2V_2^\top \nabla_{\eta\xi}\mathcal{L}(\eta^{k_u,k_l},\xi^{k_u}) \\
&- \left(V_2V_2^\top - \tau_l V_2V_2^\top \nabla_{\eta\eta}\mathcal{L}(\eta^*(\xi^{k_u}),\xi^{k_u})\right) w^{k_l} + \tau_l V_2V_2^\top \nabla_{\eta\xi}\mathcal{L}(\eta^*(\xi^{k_u}),\xi^{k_u})\|^2 \\
\leqslant (1+\gamma)&\| \left(V_2V_2^\top - \tau_l V_2V_2^\top \nabla_{\eta\eta}\mathcal{L}(\eta^{k_u,k_l},\xi^{k_u})\right)\left(\nabla_{\xi^{k_u}}\eta^{k_u,k_l} - w^{k_l}\right)\|^2 + \\
&+ 2\left(1+\frac{1}{\gamma}\right)\|\tau_l V_2V_2^\top \left(\nabla_{\eta\eta}\mathcal{L}(\eta^{k_u,k_l},\xi^{k_u}) - \nabla_{\eta\eta}\mathcal{L}\left(\eta^*(\xi^{k_u}),\xi^{k_u}\right) w^{k_l}\|^2 \right. \\
&+ 2\left(1+\frac{1}{\gamma}\right)\|\tau_l V_2V_2^\top \left(\nabla_{\eta\xi}\mathcal{L}(\eta^{k_u,k_l},\xi^{k_u}) - \nabla_{\eta\xi}\mathcal{L}\left(\eta^*(\xi^{k_u}),\xi^{k_u}\right)\|^2 \right. \\
\leqslant (1+\gamma)&(1-\tau_l\mu_l)^2\|\nabla_{\xi^{k_u}}\eta^{k_u,k_l} - w^{k_l}\|^2 \\
&+ \left(1+\frac{1}{\gamma}\right)\tau_l^2\ell_{l,2}^2\|\eta^*(\xi^{k_u}) - \eta^{k_u,k_l}\|^2\left(2+2\|w^{k_l}\|^2\right) \qquad (77)
\end{aligned}
$$

where the first inequality is derived from $\|a+b+c\|_2^2 \leqslant (1+\gamma)\|a\|_2^2 + (2+\frac{2}{\gamma})\|b\|_2^2 + (2+\frac{2}{\gamma})\|c\|_2^2$ and the second inequality is due to assumptions 1 and 2. On the one hand, $\|w^{k_l}\| \leqslant$

$\|\nabla\eta^*(\xi^{k_u})\| + \|w^{k_l} - \nabla\eta^*(\xi^{k_u})\| \leqslant \ell_{l,1}(\frac{1}{\mu_l} + \tau_l)$ is bounded according to lemma 5.2 and (76). Thus, if $\tau_l \leqslant \frac{1}{2\ell_{l,1}}$ and letting $\gamma = \tau_l\mu_l$, (77) becomes

$$\begin{aligned}\|\nabla_{\xi^{k_u}}\eta^{k_u,k_l+1} - w^{k+1}\|^2 &\leqslant (1-\tau_l\mu_l)\,\|\nabla_{\xi^{k_u}}\eta^{k_u,k_l} - w^{k_l}\|^2 \\ &\quad + \left(1 + \frac{1}{\tau_l\mu_l}\right)\tau_l^2\ell_{l,2}^2\left(\frac{4\ell_{l,1}^2}{\mu_l^2} + 3\right)\|\eta^*\left(\xi^{k_u}\right) - \eta^{k_u,k_l}\|^2 \\ &\leqslant (1-\tau_l\mu_l)\,\|\nabla_{\xi^{k_u}}\eta^{k_u,k_l} - w^{k_l}\|^2 + C_l^2\|\eta^*\left(\xi^{k_u}\right) - \eta^{k_u,k_l}\|^2 \end{aligned} \tag{78}$$

where $C_l^2 := (1 + \frac{\ell_{l,1}}{\mu_l})\ell_{l,2}^2(\frac{2}{\mu_l^2} + \frac{3}{2\ell_{l,1}^2})$.

On the other hand, we know that projected gradient descent is a contraction according to [32], i.e.

$$\|\eta^{k_u,k_l+1} - \eta^*\left(\xi^{k_u}\right)\|^2 \leqslant (1-\tau_l\mu_l)\,\|\eta^{k_u,k_l} - \eta^*\left(\xi^{k_u}\right)\|^2 \tag{79}$$

for $0 \leqslant \tau_l \leqslant \frac{1}{\ell_{l,1}}$. By induction, we have

$$\|\eta^{k_u,k_l+1} - \eta^*\left(\xi^{k_u}\right)\|^2 \leqslant (1-\tau_l\mu_l)^{k_l}\,\|\eta^{k_u} - \eta^*\left(\xi^{k_u}\right)\|^2. \tag{80}$$

Then (78) becomes

$$\|\nabla_{\xi^{k_u}}\eta^{k_u,k_l+1} - w^{k_l+1}\|^2 \leqslant (1-\tau_l\mu_l)\,\|\nabla_{\xi^{k_u}}\eta^{k_u,k_l} - w^{k_l}\|^2 + C_l^2\,(1-\tau_l\mu_l)^{k_l-1}\|\eta^{k_u} - \eta^*\left(\xi^{k_u}\right)\|^2. \tag{81}$$

Then by induction and $w^1 = \nabla_{\xi^{k_u}}\eta^{k_u,1} = 0, \eta^{k_u+1} = \eta^{k_u,K_l+1}$, we obtain that

$$\|\nabla_{\xi^{k_u}}\eta^{k_u+1} - w^{K_l+1}\|^2 \leqslant K_l\,(1-\tau_l\mu_l)^{K_l-1}\,C_l^2\|\eta^*\left(\xi^{k_u}\right) - \eta^{k_u}\|^2. \tag{82}$$

Combining (82) with (76) and setting $\tau_l \leqslant \frac{1}{2\ell_{l,1}}$, we know that

$$\|\nabla_{\xi^{k_u}}\eta^{k_u+1} - \nabla\eta^*\left(\xi^{k_u}\right)\|^2 \leqslant 2\,(1-\tau_l\mu_l)^{2K_l+2} + 2K_l\,(1-\tau_l\mu_l)^{K_l-1}\,C_l^2\|\eta^*\left(\xi^{k_u}\right) - \eta^{k_u}\|^2. \tag{83}$$

Then given τ_l and let $f(K_l) = K_l(1-\tau_l\mu_l)^{K_l-1}$, we know $\log(f(K_l)) = \log K_l + (K_l - 1)\log(1-\tau_l\mu_l)$. Taking the gradient of $\log(f(K_l))$, we get $1/K_l + \log(1-\tau_l\mu_l)$. As $\log(1-\tau_l\mu_l) < 0$, we know $\log(f(K_l))$ first increases and then decreases and thus, $\log(f(K_l))$ and $f(K_l)$ have a finite upper bound. Let us denote the upper bound of $K_l(1-\tau_l\mu_l)^{K_l-1}$ as $C_{K_l} = \mathcal{O}(1)$. Then (83) becomes

$$\|\nabla_{\xi^{k_u}}\eta^{k_u+1} - \nabla\eta^*\left(\xi^{k_u}\right)\|^2 \leqslant 2\,(1-\tau_l\mu_l)^{2K_l+2} + 2C_{K_l}C_l^2\|\eta^*\left(\xi^{k_u}\right) - \eta^{k_u}\|^2. \tag{84}$$

which yields the conclusion. □

Besides, we have the lower-level contraction and error.

Lemma 5.4 (Lower-level error). *Suppose that assumptions 1–3 hold and $\tau_l \leqslant \frac{1}{\ell_{l,1}}$, then for any $\gamma > 0$, we have*

$$\|\eta^{k_u+1} - \eta^*\left(\xi^{k_u}\right)\|^2 \leqslant (1-\tau_l\mu_l)^{K_l}\,\|\eta^{k_u} - \eta^*\left(\xi^{k_u}\right)\|^2 \tag{85a}$$

$$\|\eta^{k_u+1}-\eta^*\left(\xi^{k_u+1}\right)\|^2 \leqslant (1+\gamma)\|\eta^{k_u+1}-\eta^*\left(\xi^{k_u}\right)\|^2 + L_\eta^2\left(1+\frac{1}{\gamma}\right)\|\xi^{k_u} - \mathrm{Proj}_\Xi\left(\xi^{k_u}-\tau_u\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\|^2. \tag{85b}$$

Proof. Equation (85*a*) comes from (80) when setting $k_l = K_l$. Moreover,

$$\begin{aligned}\|\eta^{k_u+1}-\eta^*\left(\xi^{k_u+1}\right)\|^2 &= \|\eta^{k_u+1}-\eta^*\left(\xi^{k_u}\right)+\eta^*\left(\xi^{k_u}\right)-\eta^*\left(\xi^{k_u+1}\right)\|^2\\ &\overset{(a)}{\leqslant} (1+\gamma)\|\eta^{k_u+1}-\eta^*\left(\xi^{k_u}\right)\|^2 + \left(1+\frac{1}{\gamma}\right)\|\eta^*\left(\xi^{k_u}\right)-\eta^*\left(\xi^{k_u+1}\right)\|^2\\ &\overset{(b)}{\leqslant} (1+\gamma)\|\eta^{k_u+1}-\eta^*\left(\xi^{k_u}\right)\|^2 + L_\eta^2\left(1+\frac{1}{\gamma}\right)\|\xi^{k_u}-\xi^{k_u+1}\|^2\\ &= (1+\gamma)\|\eta^{k_u+1}-\eta^*\left(\xi^{k_u}\right)\|^2 + L_\eta^2\left(1+\frac{1}{\gamma}\right)\|\xi^{k_u}-\mathrm{Proj}_\Xi\left(\xi^{k_u}-\tau_u\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\|^2\end{aligned}$$

where (a) is due to $\|a+b\|_2^2 \leqslant (1+\gamma)\|a\|_2^2 + (1+\frac{1}{\gamma})\|b\|_2^2$ for any $\gamma > 0$, and (b) comes from the Lipschitz continuity of $\eta^*(\xi)$ in lemma 5.2. □

Lemma 5.5 (Upper-level error). *Under Suppose that assumptions 1–3 hold and* $\tau_l \leqslant \frac{1}{2\ell_{l,1}}$, *then it holds that*

$$\begin{aligned}u\left(\xi^{k_u+1}\right)-u\left(\xi^{k_u}\right) \leqslant& -\frac{\tau_u}{2}\|\xi^{k_u}-\mathrm{Proj}_\Xi\left(\xi^{k_u}-\nabla u\left(\xi^{k_u}\right)\right)\|^2 - \left(\frac{1}{2\tau_u}-\frac{L_u}{2}\right)\|\xi^{k_u}-\mathrm{Proj}_\Xi\left(\xi^{k_u}-\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\|^2\\ &+\tau_u\left(\ell_{u,1}\left(1+L_\eta\right)+2\ell_{u,0}C_{K_l}C_l^2\right)^2\|\eta^*\left(\xi^{k_u}\right)-\eta^{k_u}\|^2 + 2\tau_u\left(1-\tau_l\mu_l\right)^{4K_l+4}.\end{aligned}$$

Proof. According to lemma 5.2, we know $u(\xi) = \mathcal{U}(\eta^*(\xi),\xi)$ is Lipschitz smooth and

$$\nabla u\left(\xi\right) = \nabla_\xi\mathcal{U}\left(\eta^*\left(\xi\right),\xi\right) + \nabla_\xi^\top\eta^*\left(\xi\right)\nabla_\eta\mathcal{U}\left(\eta^*\left(\xi\right),\xi\right)$$

and for any ξ_1,ξ_2, we have

$$\begin{aligned}\|\nabla u(\xi_1)-\nabla u(\xi_2)\| =& \|\nabla_\xi\mathcal{U}\left(\eta^*\left(\xi_1\right),\xi_1\right)+\nabla_\xi^\top\eta^*\left(\xi_1\right)\nabla_\eta\mathcal{U}\left(\eta^*\left(\xi_1\right),\xi_1\right)-\nabla_\xi\mathcal{U}\left(\eta^*\left(\xi_2\right),\xi_2\right)\\ &-\nabla_\xi^\top\eta^*\left(\xi_2\right)\nabla_\eta\mathcal{U}\left(\eta^*\left(\xi_2\right),\xi_2\right)\|\\ \leqslant& \|\nabla_\xi\mathcal{U}\left(\eta^*\left(\xi_1\right),\xi_1\right)-\nabla_\xi\mathcal{U}\left(\eta^*\left(\xi_2\right),\xi_2\right)\| + \|\nabla_\xi^\top\eta^*\left(\xi_1\right)\|\|\nabla_\eta\mathcal{U}\left(\eta^*\left(\xi_1\right),\xi_1\right)\\ &-\nabla_\eta\mathcal{U}\left(\eta^*\left(\xi_2\right),\xi_2\right)\| + \|\nabla_\eta\mathcal{U}\left(\eta^*\left(\xi_2\right),\xi_2\right)\|\|\nabla_\xi\eta^*\left(\xi_1\right)-\nabla_\xi\eta^*(\xi_2)\|\\ \leqslant& \ell_{u,1}(\|\eta^*(\xi_1)-\eta^*(\xi_2)\|+\|\xi_1-\xi_2\|) + L_\eta\ell_{u,1}(\|\eta^*(\xi_1)-\eta^*(\xi_2)\|+\|\xi_1-\xi_2\|)\\ &+\ell_{u,0}L_{\eta\xi}\|\xi_1-\xi_2\|\\ \leqslant& (\ell_{u,1}(1+L_\eta)^2+\ell_{u,0}L_{\eta\xi})\|\xi_1-\xi_2\|.\end{aligned}$$

By denoting the smoothness constant of $u(\xi)$ as $L_u := \ell_{u,1}(1+L\eta)^2 + \ell_{u,0}L_{\eta\xi}$, we have the following expansion

$$\begin{aligned}u\left(\xi^{k_u+1}\right) &\leqslant u\left(\xi^{k_u}\right)+\left\langle\nabla u\left(\xi^{k_u}\right),\xi^{k_u+1}-\xi^{k_u}\right\rangle+\frac{L_u}{2}\|\xi^{k_u+1}-\xi^{k_u}\|^2\\ &= u\left(\xi^{k_u}\right)-\left\langle\nabla u\left(\xi^{k_u}\right),\xi^{k_u}-\mathrm{Proj}_\Xi\left(\xi^{k_u}-\tau_u\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\right\rangle+\frac{L_u}{2}\|\xi^{k_u}-\mathrm{Proj}_\Xi\left(\xi^{k_u}-\tau_u\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\|^2\\ &\overset{(a)}{=} u\left(\xi^{k_u}\right)-\frac{1}{\tau_u}\left\langle\xi^{k_u}-\mathrm{Proj}_\Xi\left(\xi^{k_u}-\tau_u\nabla u\left(\xi^{k_u}\right)\right),\xi^{k_u}-\mathrm{Proj}_\Xi\left(\xi^{k_u}-\tau_u\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\right\rangle\\ &\quad+\frac{L_u}{2}\|\xi^{k_u}-\mathrm{Proj}_\Xi\left(\xi^{k_u}-\tau_u\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\|^2\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{=} u\left(\xi^{k_u}\right) - \frac{1}{2\tau_u}\|\xi^{k_u} - \mathrm{Proj}_{\Xi}\left(\xi^{k_u} - \tau_u \nabla u\left(\xi^{k_u}\right)\right)\|^2 \\
&\quad + \frac{1}{2\tau_u}\|\mathrm{Proj}_{\Xi}\left(\xi^{k_u} - \tau_u \nabla u\left(\xi^{k_u}\right)\right) - \mathrm{Proj}_{\Xi}\left(\xi^{k_u} - \tau_u \widehat{\nabla} u\left(\xi^{k_u}\right)\right)\|^2 \\
&\quad - \left(\frac{1}{2\tau_u} - \frac{L_u}{2}\right)\|\xi^{k_u} - \mathrm{Proj}_{\Xi}(\xi^{k_u} - \tau_u \widehat{\nabla} u(\xi^{k_u}))\|^2 \\
&\overset{(c)}{\leqslant} u(\xi^{k_u}) - \frac{\tau_u}{2}\|\xi^{k_u} - \mathrm{Proj}_{\Xi}(\xi^{k_u} - \nabla u(\xi^{k_u}))\|^2 + \frac{\tau_u}{2}\|\nabla u(\xi^{k_u}) - \widehat{\nabla} u(\xi^{k_u})\|^2 \\
&\quad - \left(\frac{1}{2\tau_u} - \frac{L_u}{2}\right)\|\xi^{k_u} - \mathrm{Proj}_{\Xi}(\xi^{k_u} - \tau_u \widehat{\nabla} u(\xi^{k_u}))\|^2
\end{aligned} \tag{86}
$$

where (a) comes from $\xi^{k_u} = \mathrm{Proj}_{\Xi}(\xi^{k_u})$ and the fact that Proj_{Ξ} onto a linear equality constraint set is a linear operator, (b) is derived from $2a^\top b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$ and (c) is because Proj_{Ξ} is a linear operator and $\|\mathrm{Proj}(A) - \mathrm{Proj}(B)\| \leqslant \|A - B\|$. Besides, we can decompose the gradient bias term as follows

$$
\begin{aligned}
\|\nabla u\left(\xi^{k_u}\right) - \widehat{\nabla} u\left(\xi^{k_u}\right)\| &= \|\nabla_{\xi}\mathcal{U}\left(\eta^*\left(\xi^{k_u}\right), \xi^{k_u}\right) - \nabla\eta^*\left(\xi^{k_u}\right)^\top \nabla_{\eta}\mathcal{U}\left(\eta^*\left(\xi^{k_u}\right), \xi^{k_u}\right) \\
&\quad - \nabla_{\xi}\mathcal{U}\left(\eta^{k_u+1}, \xi^{k_u}\right) + \nabla^\top_{\xi^{k_u}}\eta^{k_u+1}\nabla_{\eta}\mathcal{U}\left(\eta^{k_u+1}, \xi^{k_u}\right)\| \\
&\leqslant \|\nabla_{\xi}\mathcal{U}\left(\eta^*\left(\xi^{k_u}\right), \xi^{k_u}\right) - \nabla_{\xi}\mathcal{U}\left(\eta^{k_u+1}, \xi^{k_u}\right)\| \\
&\quad + \|\nabla\eta^*\left(\xi^{k_u}\right)\|\|\nabla_{\eta}\mathcal{U}\left(\eta^*\left(\xi^{k_u}\right), \xi^{k_u}\right) - \nabla_{\eta}\mathcal{U}\left(\eta^{k_u+1}, \xi^{k_u}\right)\| \\
&\quad + \|\nabla_{\eta}\mathcal{U}\left(\eta^{k_u+1}, \xi^{k_u}\right)\|\|\nabla\eta^*\left(\xi^{k_u}\right) - \nabla_{\xi^{k_u}}\eta^{k_u+1}\| \\
&\leqslant \ell_{u,1}\left(1 + L_\eta\right)\|\eta^*\left(\xi^{k_u}\right) - \eta^{k_u+1}\| + \ell_{u,0}\|\nabla\eta^*\left(\xi^{k_u}\right) - \nabla_{\xi^{k_u}}\eta^{k_u+1}\| \\
&\overset{(a)}{\leqslant} \left(\ell_{u,1}\left(1 + L_\eta\right) + 2\ell_{u,0}C_{K_l}C_l^2\right)\|\eta^*(\xi^{k_u}) - \eta^{k_u}\| + 2\left(1 - \tau_l\mu_l\right)^{2K_l+2}
\end{aligned} \tag{87}
$$

where (a) comes from lower-level contraction (80). Thus, plugging (87) to (86), we get that

$$
\begin{aligned}
u\left(\xi^{k_u+1}\right) - u\left(\xi^{k_u}\right) &\leqslant -\frac{\tau_u}{2}\|\xi^{k_u} - \mathrm{Proj}_{\Xi}\left(\xi^{k_u} - \nabla u\left(\xi^{k_u}\right)\right)\|^2 - \left(\frac{1}{2\tau_u} - \frac{L_u}{2}\right)\|\xi^{k_u} - \mathrm{Proj}_{\Xi}\left(\xi^{k_u} - \widehat{\nabla} u\left(\xi^{k_u}\right)\right)\|^2 \\
&\quad + \frac{\tau_u}{2}\left(\left(\ell_{u,1}\left(1 + L_\eta\right) + 2\ell_{u,0}C_{K_l}C_l^2\right)\|\eta^*\left(\xi^{k_u}\right) - \eta^{k_u}\| + 2\left(1 - \tau_l\mu_l\right)^{2K+2}\right)^2 \\
&\leqslant -\frac{\tau_u}{2}\|\xi^{k_u} - \mathrm{Proj}_{\Xi}\left(\xi^{k_u} - \nabla u\left(\xi^{k_u}\right)\right)\|^2 - \left(\frac{1}{2\tau_u} - \frac{L_u}{2}\right)\|\xi^{k_u} - \mathrm{Proj}_{\Xi}\left(\xi^{k_u} - \widehat{\nabla} u\left(\xi^{k_u}\right)\right)\|^2 \\
&\quad + \tau_u\left(\ell_{u,1}\left(1 + L_\eta\right) + 2\ell_{u,0}C_{K_l}C_l^2\right)^2\|\eta^*\left(\xi^{k_u}\right) - \eta^{k_u}\|^2 + 2\tau_u\left(1 - \tau_l\mu_l\right)^{4K_l+4}.
\end{aligned}
$$

□

With lemmas 5.2–5.5, we restate the convergence theorem 4.3 in a more formal way and prove the theorem as follows.

Theorem 5.6. *Under assumptions 1–3, let* $\tau_l \leqslant \frac{1}{2\ell_{l,1}}, K_l = \mathcal{O}(\log K_u)$ *and* $\tau_u = \mathcal{O}(1)$ *satisfies*

$$
\tau_u \leqslant \min\left\{\frac{1}{2L_u\left(1 + 2L_\eta\right)}, \frac{\tau_l L_u \mu_l}{L_\eta\left(\left(\ell_{u,1}\left(1 + L_\eta\right) + 2\ell_{u,0}C_{K_l}C_l^2\right)^2 + 4L_u^2\right)}\right\},
$$

then the iterates of algorithm 2 satisfy

$$\frac{1}{K_u}\sum_{k_u=1}^{K_u}\|\xi^{k_u}-\mathrm{Proj}_{\Xi}\left(\xi^{k_u}-\nabla u\left(\xi^{k_u}\right)\right)\|^2=\mathcal{O}\left(\frac{1}{K_u}\right) \tag{88}$$

where $\mathcal{O}$ omits the log dependency.

Proof. We can define Lyapunov function as

$$\mathbb{V}^{k_u}=u\left(\xi^{k_u}\right)+\frac{L_u}{L_\eta}\|\eta^*\left(\xi^{k_u}\right)-\eta^{k_u}\|^2.$$

On the one hand, plugging (85*a*) to (85*b*), we get

$$\begin{aligned}\|\eta^{k_u+1}-\eta^*\left(\xi^{k_u+1}\right)\|^2\leqslant&\left(1+\gamma\right)\left(1-\tau_l\mu_l\right)\|\eta^{k_u}-\eta^*\left(\xi^{k_u}\right)\|^2\\&+L_\eta^2\left(1+\frac{1}{\gamma}\right)\|\xi^{k_u}-\mathrm{Proj}_{\Xi}\left(\xi^{k_u}-\tau_u\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\|^2.\end{aligned} \tag{89}$$

On the other hand, according to lemma 5.5 and (89), it holds that

$$\begin{aligned}\mathbb{V}^{k_u+1}-\mathbb{V}^{k_u}\leqslant&-\frac{\tau_u}{2}\|\xi^{k_u}-\mathrm{Proj}_{\Xi}\left(\xi^{k_u}-\nabla u\left(\xi^{k_u}\right)\right)\|^2-\left(\frac{1}{2\tau_u}-\frac{L_u}{2}\right)\|\xi^{k_u}-\mathrm{Proj}_{\Xi}\left(\xi^{k_u}-\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\|^2\\&+\tau_u\left(\ell_{u,1}\left(1+L_\eta\right)+2\ell_{u,0}C_{K_l}C_l^2\right)^2\|\eta^*\left(\xi^{k_u}\right)-\eta^{k_u}\|^2+2\tau_u\left(1-\tau_l\mu_l\right)^{4K+4}\\&+\frac{L_u}{L_\eta}\left[\left(1+\gamma\right)\left(1-\tau_l\mu_l\right)-1\right]\|\eta^{k_u}-\eta^*\left(\xi^{k_u}\right)\|^2+L_uL_\eta\left(1+\frac{1}{\gamma}\right)\|\xi^{k_u}-\mathrm{Proj}_{\Xi}\left(\xi^{k_u}-\tau_u\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\|^2\\\overset{(a)}{\leqslant}&-\frac{\tau_u}{2}\|\xi^{k_u}-\mathrm{Proj}_{\Xi}\left(\xi^{k_u}-\nabla u\left(\xi^{k_u}\right)\right)\|^2-\left(\frac{1}{4\tau_u}-\frac{L_u}{2}-L_uL_\eta\right)\|\xi^{k_u}-\mathrm{Proj}_{\Xi}\left(\xi^{k_u}-\tau_u\widehat{\nabla}u\left(\xi^{k_u}\right)\right)\|^2\\&-\left(\frac{\tau_lL_u\mu_l}{L_\eta}-\tau_u\left(\left(\ell_{u,1}\left(1+L_\eta\right)+2\ell_{u,0}C_{K_l}C_l^2\right)^2+4L_u^2\right)\right)\|\eta^*(\xi^{k_u})-\eta^{k_u}\|^2+2\tau_u\left(1-\tau_l\mu_l\right)^{4K_l+4}\\\overset{(b)}{\leqslant}&-\frac{\tau_u}{2}\|\xi^{k_u}-\mathrm{Proj}_{\Xi}(\xi^{k_u}-\nabla u(\xi^{k_u}))\|^2+2\tau_u\left(1-\tau_l\mu_l\right)^{4K_l+4}\end{aligned} \tag{90}$$

where (a) is earned by setting $\gamma=4L_uL_\eta\tau_u$ and (b) comes from the conditions

$$\frac{1}{4\tau_u}-\frac{L_u}{2}-L_uL_\eta\geqslant 0,\quad\text{and}\quad\frac{\tau_lL_u\mu_l}{L_\eta}-\tau_u\left(\left(\ell_{u,1}\left(1+L_\eta\right)+2\ell_{u,0}C_{K_l}C_l^2\right)^2+4L_u^2\right)\geqslant 0. \tag{91}$$

The sufficient conditions for (91) are

$$\tau_u\leqslant\min\left\{\frac{1}{2L_u\left(1+2L_\eta\right)},\frac{\tau_lL_u\mu_l}{L_\eta\left(\left(\ell_{u,1}\left(1+L_\eta\right)+2\ell_{u,0}C_{K_l}C_l^2\right)^2+4L_u^2\right)}\right\}.$$

Rearranging terms and telescoping (90) yield

$$\frac{1}{K_u}\sum_{k_u=1}^{K_u}\left\|\xi^{k_u}-\mathrm{Proj}_{\Xi}\left(\xi^{k_u}-\nabla u\left(\xi^{k_u}\right)\right)\right\|^2 \leqslant \frac{2\left(\mathbb{V}^1-\mathbb{V}^{K_u+1}\right)}{\tau_u K_u}+4\left(1-\tau_l\mu_l\right)^{4K_l+4}$$
$$\leqslant \frac{2\left(\mathbb{V}^1-\inf_{\xi} u\left(\xi\right)\right)}{\tau_u K_u}+4\left(1-\tau_l\mu_l\right)^{4K_l+4}.$$

Then by choosing $K_l=\mathcal{O}(\log(K_u))$, the convergence rate of algorithm 2 is $\mathcal{O}(\frac{1}{K_u})$. □

The above theorem guarantees the algorithm 2 converges to an ϵ stationary point given that assumptions 1–3 are satisfied. And lemma 3.4 states the convexity and Lipschitz smoothness of the lower-level objective functions $\mathcal{L}_{\mathcal{G}}(\rho,\mathbf{m};g,b)$ in (21) and shows that our problem setting satisfies the assumptions.

Since the interpolation operators I_x, I_y, I_t are linear and positive definite, to prove lemma 3.4, it is sufficient to prove the (strong) convexity and the Lipschitz smoothness of $L_{G,\gamma}:\mathbb{R}^+\times\mathbb{R}^d\to\mathbb{R},(\alpha,\boldsymbol{\beta})\mapsto\frac{\boldsymbol{\beta}^\top G\boldsymbol{\beta}}{2\alpha}+\gamma\alpha\log(\alpha)$.

Lemma 5.7. *Let G be a $d\times d$ symmetric positive definite matrix and $L_{G,\gamma}:\mathbb{R}^+\times\mathbb{R}^d\to\mathbb{R},(\alpha,\boldsymbol{\beta})\mapsto\frac{\boldsymbol{\beta}^\top G\boldsymbol{\beta}}{2\alpha}+\gamma\alpha\log(\alpha)$. For any $\gamma\geqslant 0$, $L_{G,\gamma}$ is convex in $\mathbb{R}^+\times\mathbb{R}^d$ and Lipschitz smooth in $\{\alpha\in\mathbb{R}:\alpha\geqslant\underline{c}_\rho>0\}\times\{\boldsymbol{\beta}\in\mathbb{R}^d:\|\boldsymbol{\beta}\|\leqslant\bar{c}_m\}$ ($\bar{c}_m>0$). And for any $\gamma>0$, $L_{G,\gamma}$ is strongly convex in $\{\alpha\in\mathbb{R}:\bar{c}_\rho\geqslant\alpha\geqslant\underline{c}_\rho>0\}\times\mathbb{R}^d$.*

Proof. Since G is symmetric and positive definite, we write the singular value decomposition of G as $G=U\Sigma_G U^\top$, with $UU^\top=U^\top U=I$, $\Sigma_G=\mathrm{diag}(\sigma_{G,d},\sigma_{G,d-1},\ldots,\sigma_{G,1})$, $(\sigma_{G,d}\geqslant\sigma_{G,d-1}\geqslant\cdots\geqslant\sigma_{G,1})$. And $\sigma_{G,i}, i=1,\ldots,d$ are the singular values of G. Denote $\Sigma_G^{\frac{1}{2}}:=\mathrm{diag}(\sqrt{\sigma_{G,d}},\sqrt{\sigma_{G,d-1}},\ldots,\sqrt{\sigma_{G,1}})$ and $S=\Sigma_G^{\frac{1}{2}}U^\top$. Then $G=S^\top S$ and the singular values of S are $\sigma_{S,i}=\sqrt{\sigma_{G,i}}$.

Obviously, $L_{G,\gamma}$ is twice differentiable in $\mathbb{R}^+\times\mathbb{R}^d$ and

$$\begin{aligned}\nabla^2 L_{G,\gamma}(\alpha,\boldsymbol{\beta}) &= \frac{1}{\alpha^3}\begin{bmatrix}\boldsymbol{\beta}^\top G\boldsymbol{\beta} & -\alpha\boldsymbol{\beta}^\top G^\top\\ -\alpha G\boldsymbol{\beta} & \alpha^2 G\end{bmatrix}\\ &= \frac{1}{\alpha^3}\begin{bmatrix}1 & \\ & S^\top\end{bmatrix}\begin{bmatrix}(S\boldsymbol{\beta})^\top S\boldsymbol{\beta}+\gamma\alpha^2 & -\alpha(S\boldsymbol{\beta})^\top\\ -\alpha S\boldsymbol{\beta} & \alpha^2 I\end{bmatrix}\begin{bmatrix}1 & \\ & S\end{bmatrix}\\ &= \frac{1}{\alpha^3}\begin{bmatrix}1 & \\ & S^\top\end{bmatrix}\nabla^2 L_{I,\gamma}(\alpha,S\boldsymbol{\beta})\begin{bmatrix}1 & \\ & S\end{bmatrix}.\end{aligned}\tag{92}$$

We denote the minimal and maximal singular values of $\nabla^2 L_{G,\gamma}(\alpha,\boldsymbol{\beta})$ as $\sigma_{\min}^{G,\gamma}(\alpha,\boldsymbol{\beta})$ and $\sigma_{\max}^{G,\gamma}(\alpha,\boldsymbol{\beta})$. Then we have

$$\sigma_{\min}^{G,\gamma}(\alpha,\boldsymbol{\beta})\geqslant\frac{\min(1,\sigma_{G,1})}{\alpha^3}\sigma_{\min}^{I,\gamma}(\alpha,S\boldsymbol{\beta}),\quad \sigma_{\max}^{G,\gamma}(\alpha,\boldsymbol{\beta})\leqslant\frac{\max(1,\sigma_{G,d})}{\alpha^3}\sigma_{\max}^{I,\gamma}(\alpha,S\boldsymbol{\beta}).\tag{93}$$

By computation, the eigenvalues $\lambda^{I,\gamma}(\alpha,\boldsymbol{\beta})$ of $\nabla^2 L_{I,\gamma}(\alpha,\boldsymbol{\beta})$ satisfy

$$\left(\lambda^2-\left(\|\boldsymbol{\beta}\|^2+(\gamma+1)\alpha^2\right)\lambda+\gamma\alpha^4\right)\left(\lambda-\alpha^2\right)^{d-1}=0.\tag{94}$$

Therefore $\lambda^{I,\gamma}(\alpha,\boldsymbol{\beta}) \geqslant 0$ and

$$\begin{cases} \sigma_{\min}^{I,\gamma}(\alpha,\boldsymbol{\beta}) \geqslant \dfrac{\gamma\alpha^4}{\|\boldsymbol{\beta}\|^2+(\gamma+1)\alpha^2} \geqslant 0, \\ \sigma_{\max}^{I,\gamma}(\alpha,\boldsymbol{\beta}) \leqslant \|\boldsymbol{\beta}\|^2+(\gamma+1)\alpha^2 \end{cases} \tag{95}$$

For $\gamma \geqslant 0$, $\sigma_{\min}^{G,\gamma}(\alpha,\boldsymbol{\beta}) \geqslant 0$ hold for any$\alpha > 0, \boldsymbol{\beta} \in \mathbb{R}^d$, which implies L_γ is convex. For $\gamma \geqslant 0, \alpha \geqslant \underline{c}_\rho, \|\boldsymbol{\beta}\| \leqslant \bar{c}_m$, $\sigma_{\max}^{G,\gamma}(\alpha,\boldsymbol{\beta}) \leqslant \max(1,\sigma_{G,d})(\frac{\sigma_{G,d}\bar{c}_m^2+(\gamma+1)\underline{c}_\rho^2}{\underline{c}_\rho^3})$ hold for any $\alpha \geqslant \underline{c}_\rho, \|\boldsymbol{\beta}\| \leqslant \bar{c}_m$, which implies $L_{G,\gamma}$ is Lipschitz smooth. And for $\gamma > 0, \bar{c}_\rho \geqslant \alpha \geqslant \underline{c}_\rho$, $\sigma_{\min}^{G,\gamma}(\alpha,\boldsymbol{\beta}) \geqslant \min(1,\sigma_{G,1})\min(\frac{\gamma\bar{c}_\rho}{\sigma_{G,d}\bar{c}_m^2+(\gamma+1)\bar{c}_\rho^2}, \frac{\gamma\underline{c}_\rho}{\sigma_{G,d}\bar{c}_m^2+(\gamma+1)\underline{c}_\rho^2})$.

□

6. Numerical experiments

6.1. Experiment settings

This section presents several numerical experiments to illustrate the effectiveness of our model and algorithm. We generate the data by solving the forward problem using the projected gradient descent algorithm proposed in [34] based on the FISTA algorithm [1]. In each experiment, we report the relative error versus the number of iterations for recovering the obstacle and the metric. The relative error for recovering the obstacle is

$$\sqrt{\frac{\sum_{i_x=1}^{n_x}\sum_{i_y=1}^{n_y}\left((b^{K_u})_{i_x,i_y}-\left(\widetilde{b}\right)_{i_x,i_y}\right)^2}{\sum_{i_x=1}^{n_x}\sum_{i_y=1}^{n_y}\left(\widetilde{b}\right)_{i_x,i_y}^2}}, \tag{96}$$

and the relative errors for the metrics are

$$\text{(1D)}\ \sqrt{\frac{\sum_{i_x=1}^{n_x}\left((g^{K_u})_{i_x}-(\widetilde{g})_{i_x}\right)^2}{\sum_{i_x=1}^{n_x}(\widetilde{g})_{i_x}^2}}, \qquad \text{(2D)}\ \sqrt{\frac{\sum_{i_x=1}^{n_x}\sum_{i_y=1}^{n_y}\|\,(g^{K_u})_{i_x,i_y}-(\widetilde{g})_{i_x,i_y}\,\|_F{}^2}{\sum_{i_x=1}^{n_x}\sum_{i_y=1}^{n_y}\|\,(\widetilde{g})_{i_x,i_y}\,\|_F{}^2}}, \tag{97}$$

where b^{K_u}, g^{K_u} are the numerical results after K_u upper-level updates and $\widetilde{b}, \widetilde{g}$ are the ground truth. We implement all of our numerical experiments in Matlab on a PC with an Intel(R) i7-8550U 1.80 GHz CPU and 16 GB memory.

6.2. Theoretical arguments verification

6.2.1. Algorithm convergence and obstacle unique identifiability. The first experiment aims to numerically verify the stability theorem 3.8, the unique identifiability theorem 3.9 and convergence analysis in theorem 4.3 of the bilevel algorithm with lower and upper-level constraints.

We discretize the space with $n_t = 16, n_x = n_y = 64$. Denote $p_g(x,y;\mu_x,\mu_y,\sigma_x,\sigma_y)$ as the probability density function of Gaussian distribution with mean (μ_x,μ_y) and covariance matrix $\mathrm{diag}(\sigma_x^2,\sigma_y^2)$. We feed the model with one pair of observations, i.e. $N = 1$, with $\mu_0 = p_g(\cdot,\cdot;-0.25,0,0.08,0.08)$, $\mu_1 = p_g(\cdot,\cdot;0.25,0,0.08,0.08)$ and $\gamma_I = 0.1, \gamma_T = 5$. We choose

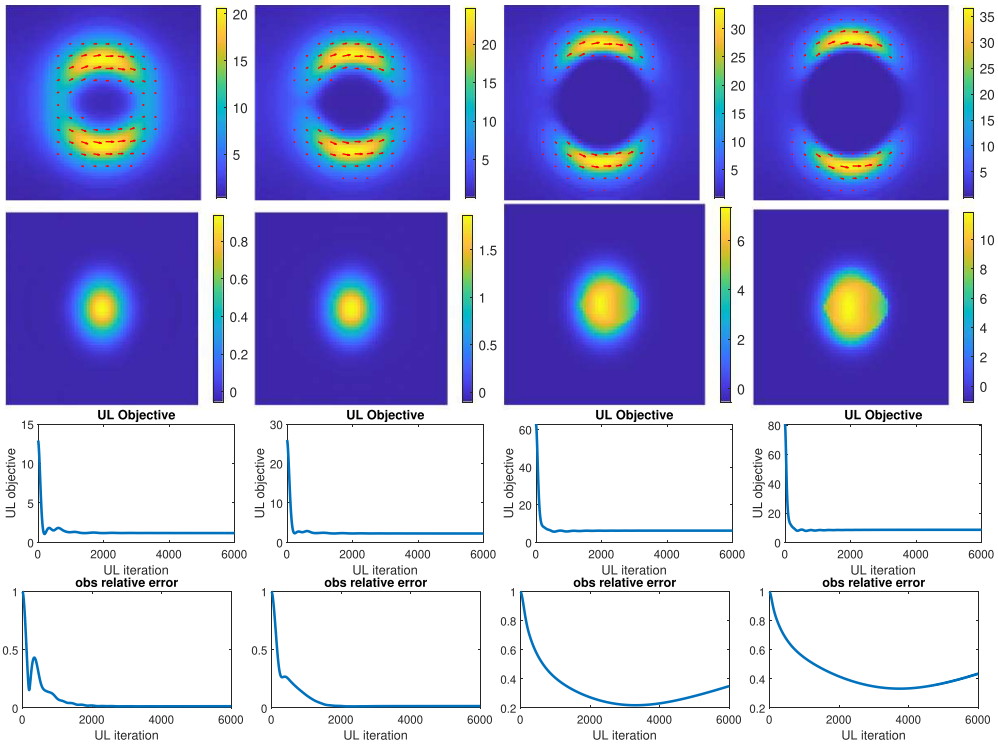

Figure 2. Convergence test of the inverse crowd motion problem. Top to bottom: the snapshot of $\widetilde{\rho}$ at $t = 0.5$, recovered b with smallest relative error, upper-level objective value versus the number of iterations, relative error of b versus the number of iterations. Left to right: $\gamma_b = 0.05, 0.1, 0.5, 1$. Reproduced with permission from [33].

Table 2. Convergence test of the inverse crowd motion problem.

γ_b	$\widetilde{\rho}(0,0,0.5)$	Upper-level objective value	Relative error (best)	Relative error (last)	Time elapsed (second)
0.05	0.7831	1.1792	0.0139	0.0148	1570.1611
0.1	0.0293	2.2504	0.0134	0.0161	1537.9703
0.5	0.0079	6.2426	0.2186	0.3500	1565.9526
1	0.0054	8.5889	0.3326	0.4354	1549.7152

the obstacle function as $b(x,y) = \gamma_b p_g(x,y;0,0,0.08,0.1)$. With different values of γ_b, the agents avoid the center of the obstacle to different degrees. Higher values of γ_b lead to lower density values at $(x,y) = (0,0)$. According to remark 3.10, low-density values in the data result in difficulties in accurately reconstructing the obstacle.

Figure 2 and table 2 compare the results with $\gamma_b = 0.05, 0.1, 0.5, 5$. For a fair comparison, we initialize the algorithm with obstacle $b^0 = \mathbf{0}$ so that the initial relative errors all start from 1 for different γ_b. We run each inner loop for 5 iterations and run the outer loop for 6000 iterations.

The first row in figure 2 plots training data $\widetilde{\rho}(\cdot,\cdot,0.5)$ and $\widetilde{\mathbf{m}}(\cdot,\cdot,0.5)$. In the first column of table 2, we report the density value $\widetilde{\rho}(\cdot,\cdot,0.5)$ at the center, reflecting the value of $\min \widetilde{\rho}$.

It is clear to see that more agents avoid the center of the obstacle as γ_b grows larger, thus the density value in the center decreases.

The third row of figure 2 presents the progression of upper-level objective values across upper-level iterations, while table 2, Column 2, details the final upper-level objective values. To enhance the precision of the upper-level objective calculation, we execute the forward solver to convergence every 10 upper-level iterations. This approach yields a refined approximation of $(\rho^*(b^{(k)}), \mathbf{m}^*(b^{(k)}))$, thereby providing a more accurate estimation of the upper-level objective values. Theorem 4.3 implies that convergence is achieved when $\min \widetilde{\rho} > 0$. Supporting this, table 2, Column 1, indicates that $\min \widetilde{\rho} > 0$ for all considered γ_b values. Furthermore, figure 2, Row 3, demonstrates numerical convergence for each γ_b selection. This verifies the algorithm convergence theorem 4.3.

We qualitatively show the numerical solutions of the obstacle in the second row of figure 2, while we report the relative error in the fourth row of figure 2 and list the best relative error and terminal step relative error in the third and fourth columns of table 2, respectively. Given that $\min \widetilde{\rho} > 0$, theorem 3.9 suggests the possibility of uniquely recovering the ground truth obstacle, up to a constant, for all γ_b values of $0.05, 0.1, 0.5$, and 1. Numerically, this unique recovery is observed for $\gamma_b = 0.05$ and 0.1. However, for higher γ_b values of 0.5 and 1, the reconstructed b does not align perfectly with the ground truth $\widetilde{b}$, as one might expect. This deviation is accounted for by remark 3.10, which discusses the robustness of the reconstruction. Specifically, when γ_b is set to 0.5 or 1, the lower bound of the data $\widetilde{\rho}$ decreases. According to remark 3.10, a smaller $\widetilde{\rho}$ lower bound leads to less robust solutions, making them more susceptible to distortions from small perturbations in the ground truth. In our experiments, since the forward solver typically produces an approximation of the exact minimizer after a finite number of iterations, the data represents a slight deviation from the ground truth. Consequently, This causes the reconstructed obstacle to differ from the exact obstacle and the discrepancy is more obvious when $\gamma_b = 0.5, 1$.

6.2.2. Improving results with multiple data. We conduct an experiment to show that multiple training data help to enhance reconstruction results for the inverse metric problem.

The example is defined on space domain $[-0.5, 0.5]$ and time domain $[0, 1]$. We discretize the space domain $[-0.5, 0.5]$ with $n_x = 64$ and the time domain $[0, 1]$ with $n_t = 16$. The ground truth metric is $\widetilde{g}(x) = 0.7 - 0.3\cos(2\pi x)$. The parameters in the forward problem are $\gamma_I = 0.01, \gamma_T = 0.5$. Then we obtain the first pair of data with $\mu_0(x) = 1.25 - 0.25\cos(4\pi x), \mu_1 = 1.25 + 0.25\cos(2\pi x)$ and the second pair with $\mu_0(x) = p_g(x; 0, 0.1), \mu_1 = 1$.

We solve the inverse problem with the first pair of data $(N = 1)$ or both data $(N = 2)$. When solving the inverse problem, we take the information on the left end $\mathcal{G}_k = \{i_x : i_x = 1\}$ as known and fix it. We choose $\mathcal{R}(g) := \frac{1}{2}\gamma_{\mathcal{R}} \int \|\nabla g(x)\|_2^2 \mathrm{d}x$ to regularize the smoothness of the metric. The discretization is therefore $\mathcal{R}_{\mathcal{G}}(g) := \frac{1}{2}\gamma_{\mathcal{R}} \Delta x \sum_{i_x=1}^{n_x-1} ((g)_{i_x+1} - (g)_{i_x})^2$.

We run the algorithm 2 for 5000 iterations with 5 iterations per each inner loop. The initialization on $i_x = 1$ is set as the true value and the initialization on other points is 0.7. Figure 3 shows the comparison of numerical results and ground truth (row 1) and the relative error of the metric versus the number of upper-level iterations (row 2). Table 3 reports the weight of regularization $\gamma_{\mathcal{R}}$, relative error, and running time of the algorithm. For one comparison, we choose no regularization $(\gamma_{\mathcal{R}} = 0)$ in the model. The results with the first data $(N = 1)$ are presented in row 1 and the results with both data $(N = 2)$ are in row 2. Then we tune the regularization parameter and report the best results with the first data in row 3 and with both data in row 4. It is easy to see that when using both data, our model captures the ground truth metric better and achieves lower relative error. It is worth noting that when using both data to solve

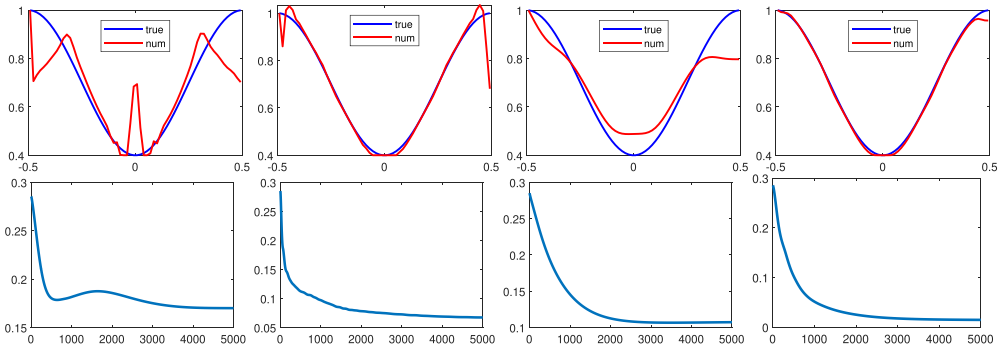

Figure 3. Improving results with multiple data. Top to bottom: comparison of numerical g and the ground truth $\widetilde{g}$, the relative error of g versus the number of iterations. Left to right: $(N=1,\gamma_{\mathcal{R}}=0)$, $(N=2,\gamma_{\mathcal{R}}=0)$, $(N=1,\gamma_{\mathcal{R}}=10^{-5})$, $(N=2,\gamma_{\mathcal{R}}=10^{-4})$.

Table 3. Improving results with multiple data.

N	$\gamma_{\mathcal{R}}$	Relative error	Time elapsed (seconds)
1	0	0.1700	60.0653
2	0	0.0673	128.5328
1	1×10^{-5}	0.1073	67.3291
2	1×10^{-4}	0.0145	118.9042

the inverse problem, our model captures the shape of the ground truth metric even without smoothness regularization. However, when using the first data, the model fails to learn the information in the center and on both ends.

6.3. *Robustness with respect to data*

6.3.1. Unknown obstacles. To test the robustness of our method for noisy input as discussed in remark 3.10, we design the following numerical experiment.

We discretize the space $[-0.5,0.5]^2$ with $n_x=n_y=64$ and choose $n_t=16$. We let the obstacle function be $b(x,y)=\begin{cases}0.5, & x<0, 0.05<y<0.1, \text{ or } x>0, -0.1<y<-0.05,\\ 0, & \text{otherwise.}\end{cases}$. Assume there is one pair of observations, with initial density $\mu_0=p_g(\cdot,\cdot;-0.3,0.3,0.1,0.1)$, preferred terminal density $\mu_1=p_g(\cdot,\cdot;0.3,-0.3,0.1,0.1)$ and $\gamma_I=0.1,\gamma_T=1$. We use the perturbed observation $\widetilde{\rho}+\gamma_n n_\rho, \widetilde{\mathbf{m}}+\gamma_n n_{\mathbf{m}}$ to solve the inverse problem, where $\gamma_n=0,0.25,0.5,0.75$ and noise $n_\rho, n_{\mathbf{m}}$ are generated by pointwise i.i.d sampling from the uniform distribution $U[-0.5,0.5]$. To avoid numerical instability caused by zero value or negative density values, we threshold the perturbed density by 0.01. All experiments initialize with the same random choice of b. Every inner loop contains 5 iterations and 5000 outer iterations have been conducted. In addition, we do not add any regularizer in this experiment. From figure 4 and table 4, we observe that with larger noise, the relative errors between numerical results and the ground truth are larger. Overall, the numerical results capture the shape of the ground truth and the algorithm converges to a close result to the ground truth $\widetilde{b}$ with reasonably low relative errors.

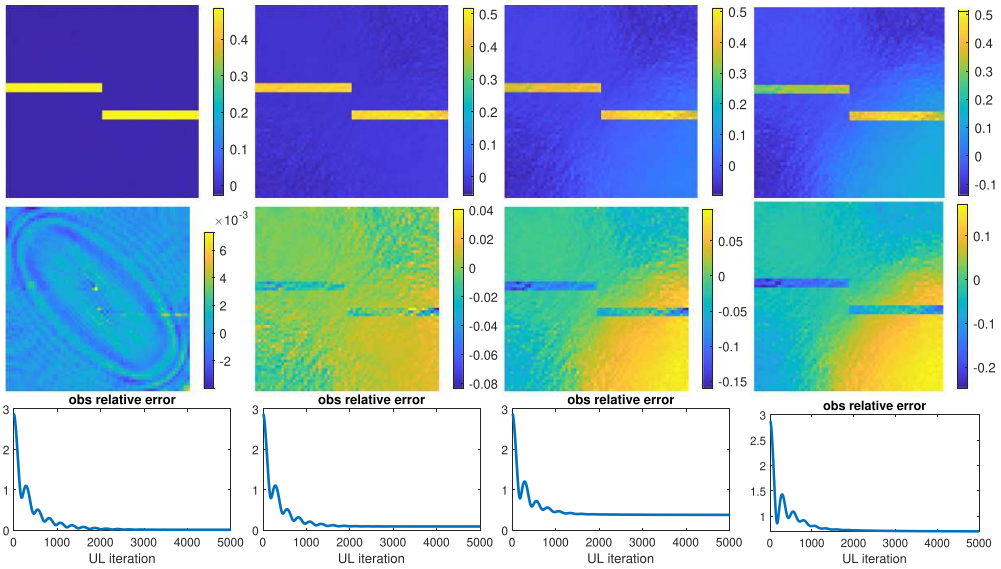

Figure 4. Robustness test of the inverse crowd motion problem. Top to bottom: numerical b, the difference between the numerical results and the ground truth $b-\widetilde{b}$, the relative error of b versus the number of iterations. Left to right: noise level $\gamma_n = 0, 0.25, 0.5, 0.75$. Reproduced with permission from [33].

Table 4. Robustness test of the inverse crowd motion problem.

γ_n	Relative error (last)	Time elapsed (second)
0	0.0081	1437.0926
0.25	0.0897	1343.8082
0.5	0.3771	1397.4578
0.75	0.7035	1379.7269

6.3.2. Unknown 1D metric. This is a 1D example on $[-0.5, 0.5] \times [0, 1]$. We discretize the space domain $[-0.5, 0.5]$ with $n_x = 64$ and the time domain $[0, 1]$ with $n_t = 16$. The ground truth metric is $\widetilde{g}(x) = 8x(x-0.375)(x+0.375)+1$. The data is obtained by taking $\mu_0(x) = p_g(x; 0, 0.1), \mu_1 = 1$ and $\gamma_I = 0.01, \gamma_T = 0.5$. We test the robustness of the model by perturbing the observation $\widetilde{\rho}, \widetilde{\mathbf{m}}$. The noises $n_\rho, n_{\mathbf{m}}$ share the same size with $\widetilde{\rho}, \widetilde{\mathbf{m}}$ and are pointwise i.i.d samples from $U[-0.5, 0.5]$. We use the perturbed data $\widetilde{\rho} + \gamma_n n_\rho, \widetilde{\mathbf{m}} + \gamma_n n_{\mathbf{m}}$ to solve the inverse problem, where $\gamma_n = 0, 0.1, 0.2, 0.3$. Row 1–2 of figure 5 illustrate the perturbed data.

When solving the inverse problem, we take the information on the left end $\mathcal{G}_k = \{i_x : i_x = 1\}$ as known and fix it. Same as section 6.2.2, we choose $\mathcal{R}(g) := \frac{1}{2}\gamma_{\mathcal{R}} \int \|\nabla g(x)\|_2^2 \mathrm{d}x$ to regularize the smoothness of the metric. The regularization weight $\gamma_{\mathcal{R}}$ takes different values for different γ_n and the values are in table 5. We run algorithm 2 for 5000 iterations with 5 iterations per each inner loop. The initialization of g takes value 1 everywhere. Figure 5 and table 5 compare the result with different γ_n.

From the comparison in figure 5 and the relative error in table 5, we observe that as the noise level increases, the recovered metric deviates more from the ground truth. However, it is crucial to highlight that, on the whole, our model adeptly captures the underlying shape of the metric with reasonable fidelity, and the associated relative error remains consistently small.

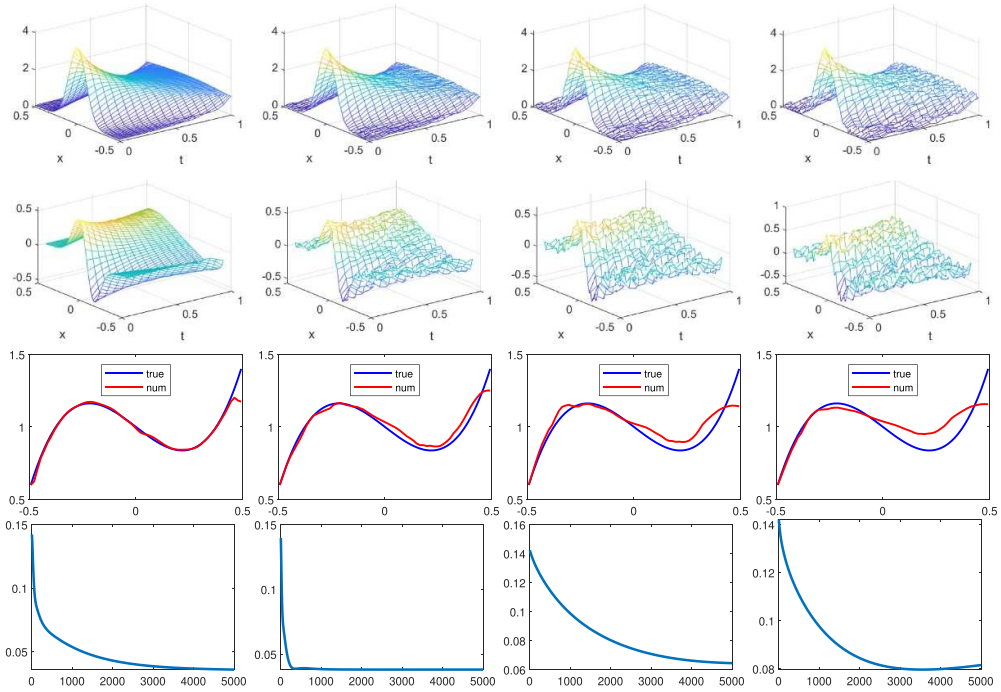

Figure 5. Robustness test of the inverse metric problem. Left to right: $\gamma_n = 0, 0.1, 0.2, 0.3$. Top to bottom: perturbed data $\widetilde{\rho} + \gamma_n n_\rho$, perturbed data $\widetilde{\mathbf{m}} + \gamma_n n_{\mathbf{m}}$, comparison of numerical g and the ground truth $\widetilde{g}$, the relative error of g versus the number of iterations.

Table 5. Robustness test of the inverse metric problem.

γ_n	$\gamma_{\mathcal{R}}$	Relative error (last)	Time elapsed (second)
0	1×10^{-5}	0.0358	63.4809
0.1	3×10^{-4}	0.0380	63.2121
0.2	1×10^{-3}	0.0645	61.5193
0.3	3×10^{-3}	0.0815	60.7215

This robust performance underscores the resilience of our model in the presence of added noise to the data.

6.4. Robustness with respect to unknowns

We present more numerical results to show that our method effectively recovers various types of obstacles and metrics.

6.4.1. Unknown obstacles. Besides the obstacle of the Gaussian type and of a 'two-bar' shape, we conduct experiments on obstacles with more irregular shapes. We plot examples of 'the segmented ring' and 'clover' in figure 6. In both experiments, only one pair of data is used to recover the unknown obstacle. The figure shows that our algorithm produces consistently

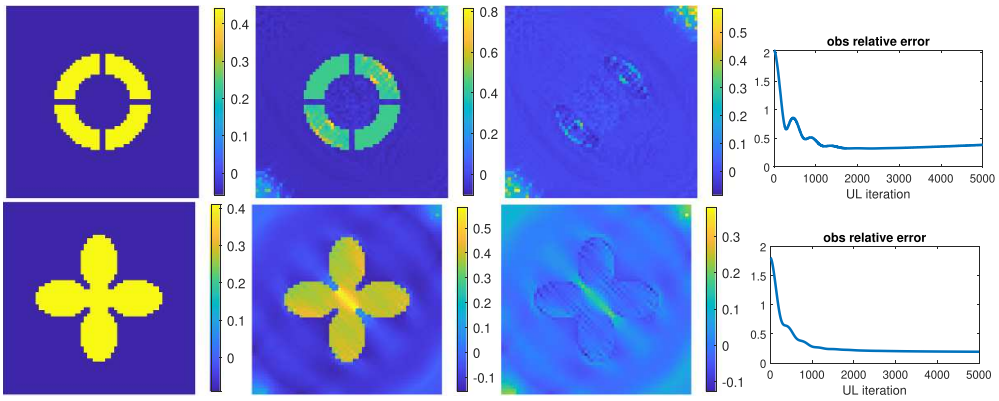

Figure 6. Robustness test of the inverse obstacle problem with respect to the obstacle. Mesh grid size: $n_t = 16, n_x = n_y = 64$. Left to right: ground truths, numerical results, the difference between ground truths and numerical results, the relative error of the obstacle versus the number of iterations. Top to bottom: relative error $= 0.3837, 0.1935$, time elapsed $= 4103$ s, 3247 s.

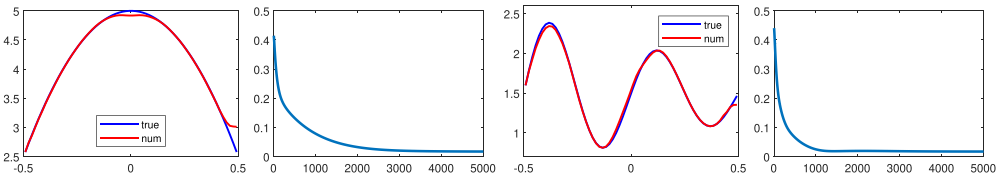

Figure 7. Robustness test of the inverse metric problem with respect to the metric. Mesh grid size: $n_t = 16, n_x = n_y = 64$. Columns 1,3: comparison of numerical g and the ground truth $\widetilde{g}$, columns 2,4: the relative error of g versus the number of iterations. Column 1,2: $\lambda = 10^{-5}$, relative error $= 0.0172$, time elapsed $= 62.8513$ s, column 3,4: $\lambda = 10^{-5}$, relative error $= 0.0172$, time elapsed $= 63.0395$ s.

good results when recovering various obstacles. Our model and algorithm recover the shape of the obstacle and achieve very low relative errors.

6.4.2. Unknown 1D metric. Apart from the experiments in sections 6.2.2 and 6.3.2, we conduct experiments on more different metrics and plot the results in figure 7. In both experiments, we use only one pair of data and the ground truth information on the left end. The figure shows that our model and algorithm consistently recover the ground truth metric and achieve low relative errors.

6.5. Unknown 2D metric

The last example is a 2D inverse metric problem on $[-0.5, 0.5]^2 \times [0, 1]$. We take $n_x = n_y = 64$ and $n_t = 16$. The ground truth metric is $\widetilde{g}(x,y) = \begin{pmatrix} g_0(x,y)+4 & g_0(x,y)+2 \\ g_0(x,y)+2 & g_0(x,y)+1 \end{pmatrix}$ with $g_0(x,y) = 0.75 + 0.5\sin(2\pi x)\cos(2\pi y - 0.5\pi)$. The data is obtained by taking $\gamma_I = 0.1, \gamma_T = 1$. We take $N = 4$, i.e. 4 observations, in this example. The initial densities are $\mu_0 = p_g(\cdot,\cdot; a_x, a_y, 0.1, 0.1)$ with $(a_x, a_y) = (-0.3, -0.3), (-0.3, 0), (-0.3, 0.3), (0, 0.3)$, and the terminal densities are

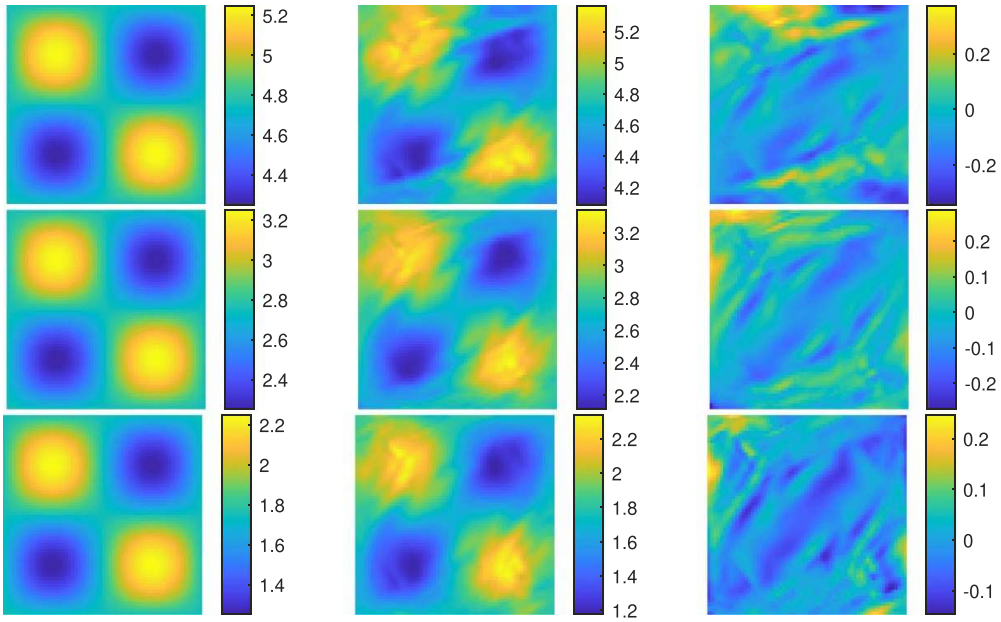

Figure 8. Solving an inverse problem with an unknown metric in 2D. Mesh grid size: $n_t = 16, n_x = n_y = 64$. Left to right: ground truths, numerical results, the difference between ground truths and numerical results. Top to bottom: g_{xx}, g_{xy}, g_{yy}. Relative error $= 0.0260$, time elapsed $= 4327.5671$ s.

$\mu_1(x,y) = p_g(\cdot,\cdot;a_x,a_y,0.1,0.1)$ with $(a_x,a_y) = (0.3,0.3),(0.3,0),(0.3,-0.3),(0,-0.3)$. We solve the inverse problem with the weights of smoothness regularizers $\gamma_{\mathcal{R}} = 10^{-4}$. The algorithm initiates from $g_{xx} = 4, g_{xy} = 2$ and $g_{yy} = 1$. Each inner loop takes 5 iterations and each outer loop takes 5000 iterations. Columns 1–3 of figure 8 shows the ground truth, the recovered metric, and the difference between the numerical result and ground truth. Our model and algorithm capture the symmetricity of the ground truth metric and achieve a relative error of value 0.0260.

7. Conclusion

In conclusion, this paper introduces a novel bilevel optimization framework to tackle inverse mean-field games for learning metrics and obstacles. We also design an alternating gradient descent algorithm to solve the proposed bilevel problems. The primary advantage of our proposed formulation is its ability to retain the convexity of the objective function and the linearity of constraints in the forward problem. Focusing on the inverse mean-field games involving unknown obstacles and metrics, we have achieved numerical stability in these setups. A significant contribution of our research is establishing unique identifiability in the inverse crowd motion model with unknown obstacles based on one pair of inputs and revealing when the solution of the bilevel problem is stable to the noisy data. Employing an alternating gradient-based optimization algorithm within our bilevel approach, we ensure its convergence and illustrate its effectiveness through comprehensive numerical experiments. These experiments serve as robust validation, underscoring the practical applicability and reliability of our algorithm in

resolving inverse problems. Our model and techniques offer a new approach to understanding and further explorations and application of inverse mean-field games.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

ORCID iDs

Jiajia Yu https://orcid.org/0000-0002-8764-8429
Quan Xiao https://orcid.org/0009-0008-8492-0037
Rongjie Lai https://orcid.org/0000-0002-3125-3321

References

[1] Beck A and Teboulle M 2009 A fast iterative shrinkage-thresholding algorithm for linear inverse problems *SIAM J. Imaging Sci.* **2** 183–202
[2] Benamou J-D and Carlier G 2014 Augmented Lagrangian methods for transport optimization, mean-field games and degenerate PDEs
[3] Benamou J-D and Carlier G 2015 Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations *J. Optim. Theory Appl.* **167** 1–26
[4] Caines P E, Huang M and Malhamé R P 2006 Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle *Commun. Inf. Syst.* **6** 221–52
[5] Chambolle A and Pock T 2011 A first-order primal-dual algorithm for convex problems with applications to imaging *J. Math. Imaging Vis.* **40** 120–45
[6] Chen T, Sun Y and Yin W 2021 Closing the gap: tighter analysis of alternating stochastic gradient methods for bilevel problems *Advances in Neural Information Processing Systems* vol 34 pp 25294–307
[7] Chow Y T, Fung S W, Liu S, Nurbekyan L and Osher S 2022 A numerical algorithm for inverse problem from partial boundary measurement arising from mean field game problem *Inverse Problems* **39** 014001
[8] Ding L, Li W, Osher S and Yin W 2022 A mean field game inverse problem *J. Sci. Comput.* **92** 7
[9] Grazzi R, Franceschi L, Pontil M and Salzo S 2020 On the iteration complexity of hypergradient computation *Int. Conf. on Machine Learning* (PMLR) pp 3748–58
[10] Guo J, Mou C, Yang X and Zhou C 2024 Decoding mean field games from population and environment observations by Gaussian processes *J. Comput. Phys.* **508** 112978
[11] Hong M, Wai H-T, Wang Z and Yang Z 2023 A two-timescale stochastic algorithm framework for bilevel optimization: complexity analysis and application to actor-critic *SIAM J. Optim.* **33** 147–80
[12] Huang M, Caines P E and Malhamé R P 2007 Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ε-Nash equilibria *IEEE Trans. Autom. Control* **52** 1560–71
[13] Imanuvilov O, Liu H and Yamamoto M 2023 Lipschitz stability for determination of states and inverse source problem for the mean field game equations (arXiv:2304.06673 [math.AP])
[14] Ji K, Yang J and Liang Y 2021 Bilevel optimization: convergence analysis and enhanced design *Int. Conference on Machine Learning* (PMLR) pp 4882–92
[15] Kachroo P, Agarwal S and Sastry S 2015 Inverse problem for non-viscous mean field control: example from traffic *IEEE Trans. Autom. Control* **61** 3412–21
[16] Klibanov M V 2023 Lipschitz stability estimate and uniqueness for a problem for the mean field games system (arXiv:2303.03928 [math.AP])
[17] Klibanov M V 2023 The mean field games system: Carleman estimates, Lipschitz stability and uniqueness (arXiv:2303.03928 [math.AP])

[18] Klibanov M V and Averboukh Y 2023 Lipschitz stability estimate and uniqueness in the retrospective analysis for the mean field games system via two Carleman estimates (arXiv:2302.10709 [math-ph])
[19] Klibanov M V and Li J 2023 The mean field games system with the lateral Cauchy data via Carleman estimates (arXiv:2303.07556 [math.AP])
[20] Klibanov M V, Li J and Liu H 2023 Holder stability and uniqueness for the mean field games system via Carleman estimates (arXiv:2304.00646 [math.AP])
[21] Klibanov M V, Li J and Yang Z 2023 Convexification for a coefficient inverse problem of mean field games (arXiv:2310.08878 [math.NA])
[22] Lasry J-M and Lions P-L 2007 Mean field games *Japan. J. Math.* **2** 229–60
[23] Lin A T, Fung S W, Li W, Nurbekyan L and Osher S J 2021 APAC-Net: alternating the population and agent control via two neural networks to solve high-dimensional stochastic mean field games *Proc. Natl Acad. Sci. USA* **118** e2024713118
[24] Liu H, Mou C and Zhang S 2022 Inverse problems for mean field games (arXiv:2205.11350 [math.OC])
[25] Liu H and Zhang S 2022 On an inverse boundary problem for mean field games (arXiv:2212.09110 [math.AP])
[26] Liu H and Zhang S 2023 Simultaneously recovering running cost and Hamiltonian in mean field games system (arXiv:2303.13096 [math.OC])
[27] Papadakis N 2015 Optimal transport for image processing *Habilitation Thesis* Université de Bordeaux
[28] Papadakis N, Peyré G and Oudet E 2014 Optimal transport with proximal splitting *SIAM J. Imaging Sci.* **7** 212–38
[29] Ren K, Soedjak N and Wang K 2023 Unique determination of cost functions in a multipopulation mean field game model (arXiv:2312.01622 [math.AP])
[30] Ruthotto L, Osher S J, Li W, Nurbekyan L and Fung S W 2020 A machine learning framework for solving high-dimensional mean field game and mean field control problems *Proc. Natl Acad. Sci.* **117** 9183–93
[31] Vicol P, Lorraine J P, Pedregosa F, Duvenaud D and Grosse R B 2022 On implicit bias in overparameterized bilevel optimization *Int. Conf. on Machine Learning* (PMLR) pp 22234–59
[32] Xiao Q, Shen H, Yin W and Chen T 2023 Alternating projected SGD for equality-constrained bilevel optimization *Int. Conf. on Artificial Intelligence and Statistics* (PMLR) pp 987–1023
[33] Yu J 2023 Numerical methods for the mean-field game and its inverse problems *PhD Thesis* Rensselaer Polytechnic Institute, New York, United States
[34] Yu J, Lai R, Li W and Osher S 2023 A fast proximal gradient method and convergence analysis for dynamic mean field planning (arXiv:2102.13260 [math.OC])