Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Automatic Control Laboratory

Distributed stochastic optimization under data-driven setting

Harleen Hanspal

Semester thesis
Spring 2018

Supervisor: Prof. Dr. John Lygeros

Advisor: Dr. Ashish Cherukuri

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

DISTRIBUTED STOCHASTIC OPTIMISATION UNDER DATA-DRIVEN SETTING

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
HANSPAL	HARLEEN

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
ZURICH, 29-06-2018	Hanspal

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Abstract

In this thesis, we solve stochastic optimization problems cooperatively in a multi-agent setting. The common goal for all agents to minimize the expected value of a function which depends on a random variable and they all have finite number of data points available with them to draw inferences about the distribution of the random variable. We adopt distributionally robust optimization (DRO) approach and characterise the uncertainity set of the DRO using Wasserstein metric as it gives desirable properties like finite sample performance guarantee, asymptotic consistency, tractability for convex problems, and is a suitable metric for data-driven setup. We reformulate the intractable DRO problem for a convex objective function into a convex problem which can be solved in polynomial time and solve the convex problem as a distributed optimization problem. For this we identify a convex-concave augmented Lagrangian function whose saddle points are in correspondence with the optimizers of the original stochastic optimisation problem and design a distributed algorithm to reach the saddle points. We derive and show using simulations the asymptotic convergence of the trajectories of the dynamics to a saddle point and hence to the optimizers of the problem. Our discussion is particularly useful for the class of objective functions which are affine in the uncertainty variable, which constitute functions used in finance for risk analysis. It also gives useful insight for much larger class of functions since piecewise affine functions frequently serve as approximations for smooth convex or concave functions.

Contents

Chapter 1

Introduction

Optimisation is the selection of the best element for a system, with regard to some criterion, from some set of available alternatives. In mathematical terms, optimisation is finding an optimal decision/optimisation variable such that certain desired objective, formulated by a function called the objective/cost or loss function, is either minimized or maximized. The set of alternatives from which this optimal decision variable is to be selected is the space spanned by the decision variables and is called the search space. This search space may be characterised by a set of constraints, defined explicitly or implicitly, in which case, the optimisation is called constrained optimisation. More formally, any optimisation problem takes the following generic form

$$\min_{x} \quad f_0(\mathbf{x}), \qquad \mathbf{x} = (x_1, ..., x_n)^T \in \mathbf{S}, \tag{1.1a}$$

$$\text{subject to} \quad f_i(\mathbf{x}) \leq 0, \qquad i = 1, ..., m \tag{1.1b}$$

$$h_i(\mathbf{x}) = 0, \qquad i = 1, ..., p \tag{1.1c}$$

Here $\mathbf{x}$ is the decision variable, the set $\mathbf{S} \subseteq \mathbb{R}^n$ is the search space, f_0 the objective function and f_i, h_i are the inequality and equality constraints respectively.
One finds many variants of the above broad definition based on whether the system is deterministic or stochastic (Stochastic Programming (**SP**) [16], Robust Optimization (**RO**)), the kinds of bounds on the search space (Semidefinite programming (**SDP**), combinatorial optimization), finite or infinite number of decision variables and constraints (Semi-Infinite Programming (**SIP**)) and also the nature of the objective function and constraints (Linear Programming (**LP**), Quadratically Constrained Quadratic Optimisation (**QCQP**), Convex optimisation [4], Non-linear optimisation, etc.) Another important variant emerges based on whether the optimal decision variable is being chosen by a single agent or by a group of agents in a distributed manner such that the optimal decision variable chosen is locally optimal for each agent and satisfies certain constraints ensuring consensus (Distributed constraint optimization (**DCOP**)).
All these variants help model different types of real world problems and a lot of work in literature can be found on each of these sub-divisions of Mathematical Optimisation. The interest is obvious since optimisation finds innumerable applications such as in business domain for portfolio optimization, supply chain management; in control engineering for designing optimal controller such as in model predictive control; in data fitting, network design and operation and almost all day to day activities like scheduling, manufacturing, inventory control, etc. in helping to make rational decisions. Many concepts in Machine Learning and statistics such as regression, Principal Component Analysis, Markov decision processes, Support Vector Machines, require setting up of an optimisation problem based on a statistical model and solving it. Typically, any optimisation problem involves two main steps, *first* setting up this optimisation problem and reformulating it into a standard form for being able to use known useful properties to solve it and *second* designing the algorithm to solve the reformulated problem. There is always an

interplay between both these steps as one solves an optimisation problem. In this project, we focus more on *multiagent*, *stochastic* optimisation in a *distributed* and *data-driven* framework.
Stochastic programming is a powerful framework for modeling optimization problems involving uncertainty. Here, the objective function (1.1a) is an expected cost

$$\mathbf{E}^{\mathbb{P}}[f_0(x,\xi)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\phi(\xi_1)\phi(\xi_2)\ldots\phi(\xi_n)f_0(x,\xi_1,\xi_2\ldots,\xi_n)d\xi_1 d\xi_2\ldots d\xi_n, \tag{1.2}$$

where the expectation is taken with respect to the distribution $\mathbb{P}$ of the continuous random variable $\xi \in \mathbb{R}^n$. In the above expression, $\phi(\xi_i)$ represents the marginal probability of ξ_i. The probability distribution $\mathbb{P}$ is usually unknown or not directly observable and hence inferred from data. In multiagent setting, this data comprises of samples collected by individual agents with or without coordination. This suffers from three challenges. *First*, even if samples were collected centrally or with complete coordination among agents, the decision variable optimising a given dataset often gives poor *out-of-sample performance* on a different dataset, even when both datasets are generated from the same distribution. This is a problem of over-fitting often faced in statistics and machine learning, where additionally the volume of dataset is much larger compared to the small finite datasets we use for optimisation problems. *Second*, for a sufficiently high accuracy, not only the exact but even the approximate solution of stochastic programs is practically intractable. For example, the approximate solution to a two-stage stochastic program with fixed recourse is #P-hard [1] even if the random variables in the problem are governed by independent uniform distributions supported on the unit hypercube $[0,1]^k$ and the objective function f_0 takes values only in the non-negitive orthant [10]. *Third*, in a distributed setting with no coordination, the data-driven solution based on only the individual agents' samples will have inferior *out-of-sample performance* due to underfitting as compared to a solution obtained collectively. A solution would be to involve a coordination strategy or sharing of data but it increases the communication and bandwidth overhead. Hence, one desires to have only as much interaction among the agents as is necessary to guarantee a solution for the problem in desired time, although with enough redundancy to ensure robustness against possible link failures.
Given the above listed issues, Distributionally Robust Optimization (**DRO**), also called the *worst case* or the *minimax* approach, is often a better and tractable alternative especially when probabilistic performance bounds and a measure of robustness against uncertainty are sought. DRO problems are robust optimization problems where the uncertainty is a probability distribution. Here, the objective function (1.1a) is the *worst-case* expected cost

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbf{E}^{\mathbb{Q}}[f_0(x;\xi)], \tag{1.3}$$

where $\mathcal{P}$ is an ambiguity set that is a family of distributions characterised by the unknown data-generating empirical distribution $\mathbb{P}$ and should contain the true distribution $\mathbb{P}$ with high probability. It produces conservative solutions in most cases but averts the problem of overfitting. However, in DRO problems, choosing a good $\mathcal{P}$ is crucial to the quality of the solution and the performance of the optimisation algorithm. Choosing a wrong ambiguity set can lead to an overly-conservative solution if set is too small or a computationally intractable problem structure if set is too large.

1.1 Literature Review

In literature, one mostly finds ambiguity sets characterised based on *generalized moment constraints* where all distributions satisfying certain moment constraints are included in the set or

[1] A function $f : \{0,1\}^* \mapsto N$ is in #P if there exists a polynomial $p : N \mapsto N$ and a polynomial-time Turing Machine M such that for every $x \in \{0,1\}^* : f(x) = |\{y \in \{0,1\}^{p(|x|)} : M(x,y) = 1\}|$. A function is #P-hard if it is at least as hard as the hardest problems in #P. [1]

based on *probability metrics* where the ambiguity set is a ball in space that contains all distributions that are close to the empirical distribution $\mathbb{P}$ in some probability distance function such as the Prokhorov metric, the Kullback-Leibler divergence or the Wasserstein metric [15]. All these probability metrics are used to represent distance functions on the space of probability distributions. While the moment based sets give tractability benefits, the metric based sets give better intuitive understanding since in metric based sets, the only parameter characterizing the size of the set is its radius. Hence, changing this parameter provides easy control over the conservativeness of the ambiguity set. However, the tractability problem as with the SP is not solved since one still has to calculate the high-dimensional integrals in the objective function [7]. Note that another possible characterisation for ambiguity sets is based on the confidence region of a goodness-of-fit (GoF) hypothesis test. However, in [3] it is explained how any ambiguity set enjoying finite-sample guarantee can be recast as the confidence region of some statistical hypothesis test and hence establishes an equivalence relation between the metric based and the goodness-of-fit based ambiguity sets.
Wasserstein ambiguity set gives apt out-of-sample performance bounds and in [8], the dependence of the set size on number of training samples is also analysed. As explained before, the *worst case expectation problem* we solve in DRO (1.3) is intractable since it has an infinite-dimensional integration over probability distributions. Therefore, tractable conic reformulations of the DRO for certain functions like the pointwise maximum of finitely many concave functions, convex functions, etc. are also derived in [8]. The computational complexity of the reformulated tractable convex functions is found to be independent of the radius of the Wasserstein ball.
In [5], distributed algorithms to find data-driven solution by a group of agents is studied for a class of objective functions such as convex-concave, convex-convex, etc. They characterise the ambiguity set based on 2-Wasserstein metric, due to which the reformulation of the DRO turns out to be convex. Carrying forward that work and using tractable reformulation for convex functions in [8] and a similar tractable reformulation derived by us in this work, we derive similar results for DRO of convex objective functions in multiagent data-driven setting with 1-Wasserstein metric. We find 1-Wasserstein metric is useful especially when minimising piecewise affine functions under uncertainty as then the DRO problem reduces to Sample Average Approximation with a compensation term. This term can be a constant as in [8] that represents the maximum stretch factor, measured with respect to the uncertainty, such that the conjugate of objective function is finite. This is intuitive as the points in the effective domain of the conjugate of a function correspond to compact sublevel sets of the function [13]. This term can also be the standard deviation of the uncertainty as in [17]. In our analysis, we define this term as the norm of the minimal subgradient of the function with respect to ξ.

1.2 Statement of contributions

In this work, we set out to solve a multiagent stochastic optimisation problem. Multiple agents cooperating with each other, by sharing their local estimates, aim to minimise the expected value of an objective function with decision variable and random variable vectors as arguments. The probability distribution of the random variable is unknown but the agents gather a finite set of samples of it to draw inferences about the distribution and find a data-driven solution of the stochastic optimization. However, the inference cannot be perfect and therefore, we adopt the Distributionally Robust Optimization approach and minimise the worst case expected value of the objective function over a neighborhood of the empirical distribution under the Wasserstein metric. This approach also gives critical performance bounds on the out-of-sample performance of our data driven solution and good generalisation results. Our first contribution is the reformulation of the DRO problem into a convex optimization with an objective function which can be expressed as the aggregate of individual objectives and whose constraints involve

consensus among neighboring agents. Therefore, its local estimates can computed by our multiple agents individually. We do a detailed analysis for the possible reformulations for the DRO problem when the objective function is convex. Next, we prove the one-to-one correspondence of the saddle points of the augmented Lagrangian function for this convex optimisation with the optimizers of the original stochastic problem. Our second contribution is the design of the saddle-point dynamics for the identified convex concave Lagrangian function which can be distributed among the agents. We prove the asymptotic convergence of its trajectories to a solution of the original stochastic optimization problem. Finally, we verify our algorithm for common convex functions, even under streaming data, in simulation.

Chapter 2

Preliminaries

This chapter introduces notation and prerequisite ideas of distributed optimisation, graph theory, convex analysis and stability of dynamical systems.

2.1 Notation

The set of real, nonnegative real, and positive integer numbers is represented by $\mathbb{R}$, $\mathbb{R}_{\geq 0}$ and $\mathbb{Z}_{\geq 1}$ and the set of extended reals as $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. $\|\cdot\|_p$ denote the p-norm on $\mathbb{R}^n$ and Euclidean norm if p is not specified. Given a norm $\|\cdot\|$ on $\mathbb{R}^m$, the dual norm is defined through $\|z\|_* := \sup_{\|\xi\|\leq 1} \langle z, \xi \rangle \sup_{x \neq 0} \frac{z^T x}{\|x\|}$. $B_\delta(x) := \{y \in \mathbb{R}^n \mid \|x - y\|_2 < \delta\}$ denotes the open ball set centered at $x \in \mathbb{R}^n$ with radius bounded by $\delta > 0$. We represent the dot product between two vectors $x, y \in \mathbb{R}^n$ as $\langle x, y \rangle := x^T y$. The support function of a set $\Xi \subseteq \mathbb{R}^m$ is defined as $\sup_{\xi \in \Xi} \langle z, \xi \rangle$. Given two sets $A_1, A_2 \subset \mathbb{R}^n, A_1 + A_2 = \{x + y \mid x \in A_1, y \in A_2\}$. Given $x, y \in \mathbb{R}^n$, x_i denotes the i-th component of x, and $x \leq y$ denotes $x_i \leq y_i$ for $i \in [n]$. For vectors $u \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$, the vector $(u; w) \in \mathbb{R}^{n+m}$ denotes their concatenation. Shorthand notations $\mathbf{0}_n = (0, \ldots, 0) \in \mathbb{R}^n$, $\mathbf{1}_n = (1, \ldots, 1) \in \mathbb{R}^n$, and $\mathsf{I}_n \in \mathbb{R}^{n \times n}$ for the identity matrix are used. For $A \in \mathbb{R}^{n_1 \times n_2}$ and $B \in \mathbb{R}^{m_1 \times m_2}$, $A \otimes B \in \mathbb{R}^{n_1 m_1 \times n_2 m_2}$ is the Kronecker product. The Cartesian product of $\{\mathcal{S}_i\}_{i=1}^n$ is denoted by $\prod_{i=1}^n \mathcal{S}_i := \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$. The interior of a set $\mathcal{S} \subset \mathbb{R}^n$ is denoted by $\mathrm{int}(\mathcal{S})$. For a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $(x, \xi) \mapsto f(x, \xi)$, $\nabla_x f : \mathbb{R}^n \to \mathbb{R}^n$ and $\nabla_\xi f : \mathbb{R}^m \to \mathbb{R}^m$ denotes the column vector of partial derivative of f with respect to the first argument and second argument respectively. The higher-order derivatives follow the convention $\nabla_{x\xi} f = \frac{\partial^2 f}{\partial x \partial \xi}$, $\nabla_{xx} f = \frac{\partial^2 f}{\partial x^2}$, and so on. Given $V : \mathcal{X} \to \mathbb{R}_{\geq 0}$, we denote the δ-sublevel set as $V^{-1}(\leq \delta) := \{x \in \mathcal{X} \mid V(x) \leq \delta\}$. We denote n times continuously differentiable function as C^n.

2.2 Graph theory

A *graph* is a triplet $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathsf{A})$, where $\mathcal{V} = [n]$ is the vertex set, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set and $\mathsf{A} \in \mathbb{R}_{\geq 0}^{n \times n}$ is the *Adjacency* matrix of $\mathcal{G}$, with the property that $a_{ij} =$ edge weight > 0 if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. A graph is *connected* if there is a path between any pair of distinct vertices. Each vertex has neighbours denoted by the set $\mathcal{N}_i = \{j \in \mathcal{V} \mid (i, j) \in \mathcal{E}\}$. If the graph is undirected, $\mathsf{A} = \mathsf{A}^T$. The *Laplacian* matrix $\mathsf{L} \in \mathbb{R}^{n \times n}$ is $\mathsf{L} = \mathsf{D} - \mathsf{A}$ where *weighted degree* matrix D is the diagonal matrix defined by $(\mathsf{D})_{ii} = w_i$, for all $i \in [n]$ and *weighted degree* of $i \in [n]$ is $w_i = \sum_{j=1}^n a_{ij}$. Clearly therefore for a connected undirected graph $\mathsf{L} = \mathsf{L}^\top$ and $\mathsf{L}\mathbf{1}_n = 0$, from which it follows that zero is a simple eigenvalue of L. The *Incidence* matrix $\mathsf{B} \in \mathbb{R}^{n \times m}$, where m is the number of edges, each assigned an arbitrary direction, in the graph

$\mathcal{G}$ is defined component-wise by

$$\mathsf{B}_{ie} = \begin{cases} +1, & \text{if node i is the source node of edge e,} \\ -1, & \text{if node i is the sink node of edge e,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

The *edge laplacian* is $\mathsf{L}_{\text{edge}} = \mathsf{B}^T\mathsf{B} \in \mathbb{R}^{m\times m}$. The Laplacian or the Incidence matrices can be used to impose consensus constraints on the vertex and edge values.

2.3 Convex analysis

A set $C \subset \mathbb{R}^n$ is *convex* if $(1-\lambda)x + \lambda y \in C$ whenever $x \in C$, $y \in C$, and $\lambda \in (0,1)$. A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *proper* if there exists $x \in \mathbb{R}^n$ such that $f(x) < +\infty$ and f does not take the value $-\infty$ anywhere in $\mathbb{R}^n$. A vector $g \in \mathbb{R}^n$ is a subgradient of $f : \mathbb{R}^n \mapsto \mathbb{R}$ at $x \in \text{dom}(f)$ if $f(z) \geq f(x) + g^T(z-x)$, for all $z \in \text{dom}(f)$. A function f is *lower semi-continuous* at x_0 if for every $\epsilon > 0$, there exists a neighborhood U of x_0 such that $f(x) \geq f(x_0) - \epsilon$ for all $x \in U$ when $f(x_0)$ is finite and $f(x) \to +\infty$ as $x \to x_0$ when $f(x_0) = +\infty$. Note that a continuous function is also lower semicontinuous but not necessarily otherwise. A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *convex* if $f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$, for all $t \in [0,1]$ and $x, y \in \text{dom}(f)$, where $\text{dom}(f)$ is the set of all points in $\mathbb{R}^n$ where f takes a finite value. If a convex function f is C^1, then it satisfies the *first-order convexity condition*, i.e., $f(y) - f(x) \geq (\nabla_x f(x))^T(y-x)$, for all $x, y \in \text{dom}(f)$. The first-order condition is equivalent to $(\nabla_y f(y) - \nabla_x f(x))^T(y-x) \geq 0$, for all $x, y \in \text{dom}(f)$. Similarly, if a convex function is C^2, then it satisfies the *second-order convexity condition*, i.e., $\nabla_{xx} f(x) \succeq 0$, for all $x \in \text{dom}(f)$. f is *strictly convex* iff it satisfies the first-order convexity condition with a strict inequality or equivalently if f is convex everywhere and $\nabla_{xx} f$ does not vanish on any non-empty open set of the domain of f. f is *strongly convex* with parameter $m > 0$ if $(\nabla_y f(y) - \nabla_x f(x))^T(y-x) \geq m\|x-y\|^2$, for all $x, y \in \text{dom}(f)$. Note that a strongly convex function is also strictly convex but the converse is not always true. A function f is concave, strictly concave, and strongly concave iff $-f$ is convex, strictly convex, and strongly convex.

A function $F : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is *convex-concave* on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^n \times \mathbb{R}^n$ if, given any point $(\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}$, $x \mapsto F(x, \tilde{y})$ is convex and $y \mapsto F(\tilde{x}, y)$ is concave. F is *closed* if for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the functions $x \mapsto F(x, y)$ and $y \mapsto -F(x, y)$ are closed. A point $(x_*, y_*) \in \mathcal{X} \times \mathcal{Y}$ is a *saddle point* of F over the set $\mathcal{X} \times \mathcal{Y}$ if $F(x_*, y) \leq F(x_*, y_*) \leq F(x, y_*)$, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. We define Saddle(F) as the set of all saddle points of F. If F is differentiable, the saddle point (x_*, y_*) is a critical point of F, i.e., $\nabla_x F(x_*, y_*) = 0$ and $\nabla_y F(x_*, y_*) = 0$. In addition, if F is C^2, then $\nabla_{xx} F(x_*, y_*) \preceq 0$ and $\nabla_{yy} F(x_*, y_*) \succeq 0$. F is *strictly convex-concave* at (x_*, y_*) if either $x \mapsto F(x, y_*)$ is strictly convex or $y \mapsto F(x_*, y)$ is strictly concave. The set of saddle points of a convex-concave function F is convex. Given a convex-concave function $F : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$, the *effective domain* of F is the product set $\mathcal{X} \times \mathcal{Y}$ defined as:

$$\begin{aligned} \mathcal{X} &:= \{x \in \mathbb{R}^n \mid F(x, y) < +\infty \text{ for all } y \in \mathbb{R}^m\}, \\ \mathcal{Y} &:= \{y \in \mathbb{R}^m \mid F(x, y) > -\infty \text{ for all } x \in \mathbb{R}^n\}. \end{aligned}$$

The sets $\mathcal{X}$, $\mathcal{Y}$ and so, $\mathcal{X} \times \mathcal{Y}$ are convex. If $\mathcal{X} \times \mathcal{Y}$ is nonempty, then this set of finite F is called *proper*. If the following equality holds

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} F(x, y) = \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} F(x, y),$$

then this common value is called the *saddle value* of F.

2.4 Distributed algorithm

In our work, we follow the following framework for designing a distributed algorithm:

(i) each agent i has the information $\mathcal{I}_i := \left\{\widehat{\Xi}_i, f, n, N, \mathbb{O}, \beta\right\}$, where $\widehat{\Xi}_i$ is the data samples gathered by i^{th} agent, f is the objective function, n is the number of agents in the network, N is the total number of data points in the complete dataset and $\mathbb{O}$ includes all the parameters associated with the nominal distribution $\mathbb{P}$ that are necessary for defining the ambiguity set as described in next section and $\beta \in (0, 1)$ is a parameter that agents agree upon beforehand and that defines the precision in the solution finally proposed,

(ii) each agent i can only exchange information with its neighbors $\mathcal{N}_i$ in the graph $\mathcal{G}$,

(iii) each agent i does not share with its neighbors any element of the dataset $\widehat{\Xi}_i$ available to it

(iv) there is no central coordinator or leader that can communicate with all agents.

Conditions (ii) and (iii) ensure that communication burden and bandwidth of the optimising strategy does not grow exponentially with the size of the network and the dataset. Since the goal of the distributed optimisation is to find the optimal values of the common decision variables with the help of all the agents, the individual agents' local optimisation is subject to a consensus constraint.

2.5 Distributionally Robust Optimisation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where Ω is the set of all possible outcomes, $\mathcal{F}$ set of events each itself a set containing zero or more outcomes and $\mathbb{P}$ the probability measure function assigning a probability to each event. Let ξ be a random variable in the corresponding Borel space $(\Omega, \mathcal{F}) := (\mathbb{R}^m, B_\sigma(\mathbb{R}^m))$. Consider the stochastic optimisation problem

$$\inf_{x \in \mathcal{X}} \mathbb{E}^{\mathbb{P}}[f(x, \xi)], \tag{2.2}$$

where the set $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ a continuous function. $\mathbb{P}$ and $\Xi \subseteq \mathbb{R}^m$ are respectively the distribution and the support of ξ. Formally, Ξ is the smallest closed set such that $\mathbb{P}(\xi \in \Xi) = 1$. $\mathbb{P}$ is unknown and therefore (2.2) cannot be solved exactly. Instead, we gather N independent and identically distributed (i.i.d.) samples $\widehat{\Xi} := \{\widehat{\xi}^k\}_{k=1}^N \subset \Xi$ of the random variable ξ. Using this dataset $\widehat{\Xi}$, we try to find a feasible decision $\widehat{x}_N \in \mathcal{X}$ with lowest possible *out-of-sample performance* i.e., minimum $\mathbb{E}^{\mathbb{P}}[f(\widehat{x}_N, \xi)]$ where ξ is an unseen sample chosen independently of the training dataset. Since the samples are i.i.d., the probability over the entire uncertainity sample space can be expressed as the product of the probabilities over the individual uncertainity samples, i.e., $\mathbb{P}^N := \prod_{i=1}^N \mathbb{P}_i$ and this probability is supported on $\Xi^N := \prod_{i=1}^N \Xi_i$. As $\mathbb{P}$ is unknown, we seek performance guarantee of the following form:

$$\mathbb{P}^N(\mathbb{E}^{\mathbb{P}}[f(\widehat{x}_N, \xi)] \leq \widehat{J}_N) \geq 1 - \beta, \tag{2.3}$$

where $\widehat{J}_N$ and $1-\beta$ is called the *certificate* and the *reliability* of this guarantee respectively. In [8], it is shown that for an appropriate Wasserstein ambiguity set (defined below), $\widehat{P}_N := \mathcal{B}_{\epsilon_N(\beta)}(\widehat{\mathbb{P}}_N)$ centered around the empirical distribution $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{k=1}^N \delta_{\widehat{\xi}^k}$, a certificate $\widehat{J}_N$ defined as

$$\widehat{J}_N := \inf_{x \in \mathcal{X}} \sup_{\mathbb{Q} \in \widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}[f(x, \xi)], \tag{2.4}$$

gives a confidence bound of the type (2.3) where $\widehat{x}_N$ is the solution of (2.4). The guarantee of this type is called the *finite-sample guarantee*. Both the data-driven solution $\widehat{x}_N$ and $\widehat{J}_N$ we obtain depends on how strong a probablistic guarantee $\beta \in (0,1)$ we want. The *asymptotic consistency* and *tractability* for a well designed Wasserstein ambiguity sets is also proven in [8].

Next, we state the design of the Wasserstein ambiguity set that provides the required guarantee given the number of samples N, the reliability $1-\beta$, and the characteristics of the distribution $\mathbb{P}$.

Definition 2.5.1. *The* Wasserstein metric*, also called the Earth mover's distance,* $d_W : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \to \mathbb{R}_{\geq 0}$ *is*

$$d_W(\mathbb{Q}_1, \mathbb{Q}_2) = \Big(\inf\Big\{ \int_{\Xi^2} \|\xi_1 - \xi_2\| \, \Pi(d\xi_1, d\xi_2) \Big| \Pi \in \mathcal{H}(\mathbb{Q}_1, \mathbb{Q}_2) \Big\}\Big), \tag{2.5}$$

where $\mathcal{H}(\mathbb{Q}_1, \mathbb{Q}_2)$ *is the set of all distributions on* $\Xi \times \Xi$ *with marginals* $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{M}(\Xi)$ *and* $\mathcal{M}(\Xi)$ *is the space of probability distributions* $\mathbb{Q}$ *supported on* Ξ *with finite first moment, i.e.* $\mathbb{E}^{\mathbb{Q}}[\|\xi\|] = \int_{\Xi} \|\xi\| \, \mathbb{Q}(d\xi) < +\infty$.

The norm encodes the cost of moving things (probability mass) from a given configuration (a distribution, say $\mathbb{Q}_1$) to another configuration (a distribution, say $\mathbb{Q}_2$). One can appreciate that the norm in (2.5) can be an arbitrary p-norm and hence gives flexibility in defining costs between different distributions.
In this metric, let $\mathcal{B}_\epsilon(\widehat{\mathbb{P}}_N)$ be the set of distributions that are ϵ-close to $\widehat{\mathbb{P}}_N$, where $\epsilon \geq 0$, i.e.,

$$\mathcal{B}_\epsilon(\widehat{\mathbb{P}}_N) := \{\mathbb{Q} \in \mathcal{M}(\Xi) \mid d_{W_2}(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq \epsilon\}. \tag{2.6}$$

We make a necessary assumption that the generating distribution $\mathbb{P}$ is a *light-tailed distribution*, i.e., there exists a $a > 1$ such that

$$b := \mathbb{E}^{\mathbb{P}}[exp(\|\xi\|^a)] < \infty. \tag{2.7}$$

Then [8, Theorem 3.4], gives the following result for the light-tailed distribution

$$\mathbb{P}^N(d_W(\mathbb{P}, \widehat{\mathbb{P}}_N) \geq \epsilon) \leq \begin{cases} c_1 e^{-c_2 N \epsilon^{\max\{m,2\}}}, & \text{if } \epsilon \leq 1, \\ c_1 e^{-c_2 N \epsilon^a}, & \text{if } \epsilon > 1, \end{cases} \tag{2.8}$$

for all $N \geq 1, m \neq 2$ and $\epsilon \geq 0$, where $c_1, c_2 \in \mathbb{R}_{>0}$ are functions of a, b (defined for a light-tailed distribution (2.7)) and a constant m. Since we want the ball $\mathcal{B}_\epsilon(\widehat{\mathbb{P}}_N)$ to contain $\mathbb{P}$ with confidence $1-\beta$, the right-hand side in (2.8) should be equal to β. Therefore, equating right-hand side in (2.8) to β, we get the radius $\epsilon_N(\beta)$ of the Wasserstein ambiguity set as

$$\epsilon_N(\beta) := \begin{cases} \left(\frac{\log(c_1 \beta^{-1})}{c_2 N}\right)^{1/\max\{m,2\}}, & \text{if } N \geq \frac{\log(c_1\beta^{-1})}{c_2}, \\ \left(\frac{\log(c_1 \beta^{-1})}{c_2 N}\right)^{1/a}, & \text{if } N < \frac{\log(c_1\beta^{-1})}{c_2}. \end{cases} \tag{2.9}$$

As a direct consequence of (2.8) above, i.e., $\mathbb{P}^N(d_W(\mathbb{P}, \widehat{\mathbb{P}}_N) \geq \epsilon) \leq \beta$, we get that

$$\mathbb{P}^N\big(\mathbb{E}^{\mathbb{P}}[f(\widehat{x}_N, \xi)] \leq \inf_{x \in \mathcal{X}} \sup_{\mathbb{Q} \in \widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}[f(x, \xi)]\big) \geq 1 - \beta$$

which is the desired *finite-sample guarantee*. As an analogy, we can analyse the closeness of $\mathbb{P}$ and $\widehat{\mathbb{P}}_N$ as a statistical hypothesis test with the $d_W(\mathbb{P}, \widehat{\mathbb{P}}_N)$ is the *test statistic* and β is the *significance parameter* with respect to $\widehat{\mathbb{P}}_N$ and ambiguity set $\widehat{\mathcal{P}}_N$ the threshold for closeness of the two distributions.

2.6 Stability of Dynamical Systems

In our work, we use LaSalle Invariance Principle for proving the asymptotic convergence of our distributed algorithm.

Theorem 2.6.1. *(LaSalle Invariance Principle [11, Theorem 4.4]): Let $\Omega \subset D$ be a compact set that is positively invariant with respect to $\dot{x} = f(x)$. Let $V : D \mapsto R$ be a C^1 function such that $\nabla_x V(x)^T f(x) \leq 0$ in Ω. Let E be the set of all points in Ω where $\nabla_x V(x)^T f(x) = 0$, then every solution starting in Ω approaches the largest invariant set in E as $t \mapsto \infty$.*

Chapter 3

Problem Statement

Consider a network of $n \in \mathbb{Z}_{\geq 1}$ agents. We model the interaction among the agents with an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V} = [n]$ is the set of vertices representing the agents, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges, and $\mathcal{A}$ the adjacency matrix of $\mathcal{G}$ such that its elements $a_{ij} > 0$ if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. If $(i, j) \in \mathcal{E}$, i and j are neighbours and can exchange information over the network. Let $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}, (x, \xi) \mapsto f(x, \xi)$ be a C^1 convex function. We assume that all agents know f and aim to solve the following stochastic optimisation problem,

$$\inf_{x \in \mathbb{R}^d} \mathbb{E}^{\mathbb{P}}[f(x, \xi)], \tag{3.1}$$

where x is the decision variable and ξ is a random variable with support $\Xi \subseteq \mathbb{R}^m$ and distribution $\mathbb{P}$. As explained in Section 2.5, $\mathbb{P}$ is unknown to the agents and hence (3.1) cannot be solved exactly. However, we assume that, to infer about $\mathbb{P}$, each agent collects finite number of independent and identically distributed (i.i.d.) samples of the uncertainty ξ. With these samples, $\widehat{\Xi}_i := \{\widehat{\xi}^k\}_{k=1}^{N_i} \subset \Xi$, collected by each agent $i \in [n]$, such that $\widehat{\Xi}_i \neq \emptyset$ and $\widehat{\Xi}_N = \cup_{i=1} \widehat{\Xi}_i$, the agents collectively find a data driven solution $\widehat{x}_N \in \mathbb{R}^d$ of (3.1) by solving the following Distributionally Robust Optimisation (DRO) problem

$$\widehat{J}_N := \inf_{x \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathcal{B}_{\epsilon_N(\beta)}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[f(x, \xi)]. \tag{3.2}$$

The data driven solution, $\widehat{x}_N$, satisfies the following finite-sample guarantee

$$\mathbb{P}^N(\mathbb{E}^{\mathbb{P}}[f(\widehat{x}_N, \xi)] \leq \widehat{J}_N) \geq 1 - \beta. \tag{3.3}$$

As explained in Section 2.5, in DRO, our objective is to find $\widehat{x}_N$ with lowest *out-of-sample performance*, i.e., minimum $\mathbb{E}^{\mathbb{P}}[f(\widehat{x}_N, \xi)]$ where ξ is an unseen sample. While finding the absolute minimum is not possible since $\mathbb{P}$ is unknown, the data-driven solution $\widehat{x}_N$ can provide *finite-sample guarantee* of the type (3.3) when for a suitable size of the ambiguity set, the certificate $\widehat{J}_N$ upper bounds the out-of-sample performance of $\widehat{x}_N$ with $1 - \beta$ confidence. Therefore, now our goal changes to finding the lowest possible certificate with the desired high confidence bound. There is a direct interplay between this confidence bound and the size of the ambiguity set, as explained in section 2.5. In most cases, one should also further refine the size of the ambiguity set, designed using the theoretical results in section 2.5, by using numerical estimation techniques such as cross-validation, bootstrapping or by using goodness-of-fit tests. Assuming that the ambiguity set has been chosen, in the subsequent chapters we solve (3.2) for the chosen $\epsilon_N(\beta)$. Additionally, note that the DRO (3.2) is intractable, as it has an infinite-dimensional expectation over probability distributions in the ambiguity set. Therefore, in the next section, we give its tractable reformulation.

3.1 Tractable reformulation of DRO

In this section, we first state and analyse reformulation of (3.2) given in [8] and give the class of objective functions which can be optimised using it. Later, we do a similar analysis for another reformulation.

Theorem 3.1.1. *(General tractable reformulation of DRO [8, Theorem 6.3]): For a proper, lower semicontinuous objective function f and a closed, convex support Ξ for the uncertainty, the following holds for all $x \in \mathbb{R}^d$*

$$\sup_{\mathbb{Q}\in\mathcal{B}_{\epsilon_N(\beta)}} \mathbb{E}^{\mathbb{Q}}[f(x,\xi)] = \inf_{\lambda\geq 0}\Big\{\lambda\epsilon_N(\beta) + \frac{1}{N}\sum_{k=1}^{N}\sup_{\xi\in\Xi}\Big(f(x,\xi) - \lambda\left\|\xi - \widehat{\xi}_k\right\|\Big)\Big\}. \tag{3.4}$$

Using strong duality results for conic linear programs from [14], one can verify that the above holds true for non-convex functions too. Solving further the above expression for the proper, convex and continuous objective function, we get the following inequality [8, Theorem 6.3]:

$$\sup_{\xi\in\Xi}\Big(f(x,\xi) - \lambda\left\|\xi - \widehat{\xi}_k\right\|\Big)\Big\} \leq \begin{cases} f(x,\widehat{\xi}_k), & \text{if } \sup\{\|\theta\| : f^*(x,\theta) < \infty\} \leq \lambda, \\ \infty, & \text{otherwise }, \end{cases} \tag{3.5}$$

where for a given x, the function $\theta \mapsto f^*(x,\theta)$ is the conjugate of $\xi \mapsto f(x,\xi)$ given by $f^*(x,\theta) = \sup_{\xi\in\Xi}\langle\theta,\xi\rangle - f(x,\xi)$. Using the above inequality, we state the following theorem.

Theorem 3.1.2. *(Tractable reformulation of (3.2) [8, Theorem 6.3]): For a C^1 convex objective function $(x,\xi) \mapsto f(x,\xi)$, the DRO problem (3.2) is smaller or equal to*

$$\inf_{x\in\mathbb{R}^d}\ \kappa(x)\epsilon_N(\beta) + \frac{1}{N}\sum_{k=1}^{N} f(x,\widehat{\xi}_k), \tag{3.6}$$

where $x \mapsto \kappa(x) := \sup\{\|\theta\| : f^*(x,\theta) < \infty\}$ for all $x \in \mathbb{R}^d$. If $\Xi = \mathbb{R}^m$, then (3.2) coincides with (3.6).

Remark 3.1.3. *$\kappa(x)$ is the smallest radius of a ball enclosing the effective domain of the conjugate function $\xi \mapsto f^*(x,\xi)$.*

Lemma 3.1.4. *Consider a function $(x,\xi) \mapsto f(x,\xi)$ and $\kappa(x) := \sup\{\|\theta\| : f^*(x,\theta) < \infty\}$ then for all $x \in \mathbb{R}^n$ and $\xi \in \mathbb{R}^m$, the following holds*

(i) If $\xi \mapsto f(x,\xi)$ is L_u-Lipschitz continuous, i.e., there exists a $\xi' \in \mathbb{R}^m$ such that $f(x,\xi) - f(x,\xi') \leq L_u\left\|\xi - \xi'\right\|$, for all $\xi \in \mathbb{R}^m$, then $\kappa(x) \leq L_u$.

(ii) If there exists a $\theta \in \mathbb{R}^m$ with $\|\theta\| = L_l$ and $\xi' \in \mathbb{R}^m$ such that $f(x,\xi) - f(x,\xi') \geq \langle\theta,\xi - \xi'\rangle$, for all $\xi \in \mathbb{R}^m$, then $\kappa(x) \geq L_l$.

Proof. For (i), note that

$$f^*(x,\theta) = \sup_{\xi\in\mathbb{R}^m}\langle\theta,\xi\rangle - f(x,\xi) \geq \sup_{\xi\in\mathbb{R}^m}\langle\theta,\xi\rangle - L_u\left\|\xi - \xi'\right\| - f(x,\xi') \tag{3.7}$$

$$= \sup_{\xi\in\mathbb{R}^m}\inf_{\|z\|\leq L_u}\langle\theta,\xi\rangle - \langle z,\xi - \xi'\rangle - f(x,\xi') \tag{3.8}$$

$$\geq \begin{cases} \langle\theta,\xi'\rangle - f(x,\xi'), & \text{if } \|\theta\| \leq L_u, \\ \infty, & \text{otherwise.} \end{cases} \tag{3.9}$$

In the above, (3.7) comes from the definition of Lipschitz continuity and conjugacy and (3.8) comes from the definition of dual norm. Clearly, $f^*(x,\theta) < \infty$ needs $\|\theta\| \leq L_u$ and therefore, $\kappa(x) \leq L_u$.
For (ii), using the definition of conjugacy and the assumption in (ii), we get the following

$$f^*(x,\theta) = \sup_{\xi\in\mathbb{R}^m} \langle\theta,\xi\rangle - f(x,\xi) \leq \sup_{\xi\in\mathbb{R}^m} \langle\theta,\xi\rangle - \langle z,\xi-\xi^{'}\rangle - f(x,\xi^{'}) \tag{3.10}$$

$$= \sup_{\xi\in\mathbb{R}^m} \langle\theta-z,\xi\rangle - \langle z,\xi^{'}\rangle - f(x,\xi^{'}) \tag{3.11}$$

From the above analysis, we can see that $f^*(x,\theta)$ is finite when $\theta = z$ and ∞ otherwise. For the former case, $f^*(x,\theta) \leq \langle\theta,\xi\rangle - f(x,\xi^{'})$ and θ belongs to the effective domain of f^*. Therefore, $\kappa \geq \|\theta\| = L_l$. ■

Remark 3.1.5. *Note that for any $x \in \mathbb{R}^d$, $\kappa(x)$ is unbounded if $f^*(x,\theta)$ is bounded for all $\theta \in \mathbb{R}^m$. From the definition of conjugate function, bounded $f^*(x,\theta)$ implies that $\langle\theta,\xi\rangle - f(x,\xi)$, which is a concave function for a convex f, must be bounded from above. The later condition is true for all proper convex functions $\xi \mapsto f(x,\xi)$. However, when f is affine in ξ, $\kappa(x)$ might be bounded.* •

Using the bounds in Lemma 3.1.4, if $\xi \mapsto f(x,\xi)$ is L_u-Lipschitz continuous, we can replace $\kappa(x)$ with a convex $L_u(x)$ such that (3.2) becomes convex. Note that we define L_u as the sharpest Lipschitz constant of the map $\xi \mapsto f(x,\xi)$, i.e., L_u is the smallest real number for which the inequality in (i) holds. However, replacing $\kappa(x)$ with $L_u(x)$ makes the solution (3.6) conservative and it is no longer exact even when $\Xi = \mathbb{R}^m$. Therefore, the certificate $\widehat{J}_N$ we get is higher than what we would get with $\kappa(x)$ in (3.6).
If the Lipschitz constant of $\xi \mapsto f(x,\xi)$ is independent of x, $\kappa(x)$ is equal to L_u, the uniform Lipschitz constant of $\xi \mapsto f(x,\xi)$, and (3.2) reduces to the following convex problem

$$\inf_{x\in\mathbb{R}^d} \kappa\epsilon_N(\beta) + \frac{1}{N}\sum_{k=1}^{N} f(x,\widehat{\xi}_k). \tag{3.12}$$

Remark 3.1.6. *The map $\xi \mapsto f(x,\xi)$ is uniformly Lipschitz in ξ for the following convex functions:*

(i) The function $f : \mathcal{X}\times\Xi \mapsto \mathbb{R}$ defined by $f(x,\xi) = \max\{x^p - \xi, 0\}$, where p is an even, non-negative integer.
Such functions are commonly used in solving ambiguity-averse and risk-neutral Newsvendor profit optimisation problems [2, 12] with demand uncertainty being modeled by ξ.

(ii) The function $f : \mathcal{X}\times\Xi \mapsto \mathbb{R}$ defined as $f(x,\xi) = f_1(x) + f_2(\xi)$, where $f_2 : \Xi \mapsto \mathbb{R}$ is Lipschitz continuous.

(iii) Any piecewise linear lower semi-continuous function.

(iv) Any continuous function $f : \mathcal{X}\times\Xi \mapsto \mathbb{R}$ for which the domain is bounded.

•

Below, we give another reformulation for the DRO problem (3.2).

Proposition 3.1.7. *(Tractable reformulation of DRO for a convex objective function affine in uncertainty): For a continuous convex objective function f affine in uncertainty, the following holds for all $x \in \mathbb{R}^d$*

$$\sup_{\mathbb{Q}\in\mathcal{B}_{\epsilon_N(\beta)}} \mathbb{E}^{\mathbb{Q}}[f(x,\xi)] \leq \|g(x)\| \frac{1}{N}\sum_{k=1}^{N} \epsilon_N(\beta) + f(x,\widehat{\xi}_k), \tag{3.13}$$

where $g(x) \in \mathbb{R}^m$ is a subgradient of $\xi \mapsto f(x,\xi)$. In addition, if the function f is C^1 then $g(x) = \nabla_\xi f(x,\xi)$.

Proof. From (3.4), we get

$$\sup_{\mathbb{Q}\in\mathcal{B}_{\epsilon_N(\beta)}} \mathbb{E}^{\mathbb{Q}}[f(x,\xi)] = \inf_{\lambda\geq 0} \sup_{\{\xi_k\}\in\Xi^N} \lambda\epsilon_N(\beta) + \frac{1}{N}\sum_{k=1}^{N}\Big(f(x,\xi_k) - \lambda\left\|\xi_k - \widehat{\xi}_k\right\|\Big). \tag{3.14}$$

Here, we have extracted out the set of supremums on the right-hand side. From the first-order convexity condition, for each k, we have

$$f(x,\xi^{'}) \geq f(x,\xi_k) + g(x)^T(\xi^{'} - \xi_k), \tag{3.15}$$

for all $\xi^{'} \in \mathbb{R}^m$. Replacing $\xi^{'}$ with $\widehat{\xi}_k$ in the above inequality and using the bound in (3.14), we get

$$\sup_{\mathbb{Q}\in\mathcal{B}_{\epsilon_N(\beta)}} \mathbb{E}^{\mathbb{Q}}[f(x,\xi)] \tag{3.16}$$

$$\leq \inf_{\lambda\geq 0} \sup_{\{\xi_k\}\in\Xi^N} \lambda\epsilon_N(\beta) + \frac{1}{N}\sum_{k=1}^{N}\Big(f(x,\widehat{\xi}_k) + g(x)^T(\xi_k - \widehat{\xi}_k) - \lambda\left\|\xi_k - \widehat{\xi}_k\right\|\Big) \tag{3.17}$$

$$\leq \inf_{\lambda\geq 0} \sup_{\{\xi_k\}\in\Xi^N} \lambda\epsilon_N(\beta) + \frac{1}{N}\sum_{k=1}^{N}\Big(f(x,\widehat{\xi}_k) + (\|g(x)\| - \lambda)\left\|\xi_k - \widehat{\xi}_k\right\|\Big) \tag{3.18}$$

$$= \begin{cases} \frac{1}{N}\sum_{k=1}^{N}\big(\|g(x)\|\,\epsilon_N(\beta) + f(x,\widehat{\xi}_k)\big), & \text{if } \|g(x)\| = \lambda, \\ \infty, & \text{otherwise.} \end{cases} \tag{3.19}$$

Implication (3.18) comes from Cauchy-Schwartz inequality. We establish implication (3.19) by verifying it for both possible cases for $\|g(x,\xi_k)\|$ and λ. First, if for some x, $\|g(x)\| > \lambda$, then we can keep increasing ξ_k and λ indefinitely to get $+\infty$. Similarly, if $\|g(x)\| \leq \lambda$, we get a finite value (3.19)(a) for (3.18) only when $\|g(x)\| \leq \lambda$. ■

Note that, the above reformulation or equivalently reformulation (3.6) where we replace $\kappa(x)$ with a convex $L_u(x)$ always gives a lower bound than (3.12) where we replace $\kappa(x)$ with a constant L_u. For example, taking $f(x,\xi) = \xi x^p$, where $x \in \mathbb{R}, \xi \in \mathbb{R}$ and p is a positive, even integer, we get the following bounds on the DRO problem (3.2) using the two reformulations.
Using reformulation (3.12), we get (3.2) is equal to ∞ if $\mathcal{X}$ is unbounded and less than or equal to the following otherwise

$$\inf_{x\in\mathcal{X}} |M|^p\epsilon_N(\beta) + \frac{1}{N}\sum_{k=1}^{N} f(x,\widehat{\xi}_k), \tag{3.20}$$

where $M > 0$ such that $|x| \leq |M|$, for all $x \in \mathcal{X}$.
Using reformulation (3.13) or with reformulation (3.6), we get (3.2) is less than or equal to

$$\inf_{x\in\mathcal{X}} |x|^p\epsilon_N(\beta) + \frac{1}{N}\sum_{k=1}^{N} f(x,\widehat{\xi}_k), \tag{3.21}$$

which is always finite for a proper objective function and always less than or equal to (3.20). However, note that (3.13) is not convex, refraining us from doing convex analysis on this reformulation. Therefore, in the reformulation (3.13), we replace g with a function $\tilde{g} : \mathbb{R}^d \mapsto \mathbb{R}^m$ defined as $\tilde{g}_i := \max\{g_i, 0\}$, for all $i \in [d]$. Note that $0 \leq g \leq \tilde{g} \implies \|g\| \leq \|\tilde{g}\|$ and therefore, we can

replace it in (3.13) such that it still upper bounds the DRO problem. Also, we prove below that if $x \mapsto g(x)$ is convex then $x \mapsto \|\tilde{g}(x)\|$ is also convex. Therefore, with the above replacement and the assumption that $x \mapsto g(x)$ is convex, we get the following modified reformulation

$$\|\tilde{g}\| \, \epsilon_N(\beta) + \frac{1}{N} \sum_{k=1}^{N} f(x, \widehat{\xi}_k). \tag{3.22}$$

Now, let a function $l : \mathbb{R}^d \mapsto \mathbb{R}$ be defined as $l(z) := \|\tilde{g}(z)\|$, then for all $\lambda \in [0, 1]$ and $z, y \in \mathbb{R}^d$, we have

$$\begin{aligned} l(\lambda z + (1-\lambda)y) &= \|\tilde{g}(\lambda z + (1-\lambda)y)\|, & (3.23)\\ &\leq \|\lambda \tilde{g}(z) + (1-\lambda)\tilde{g}(y)\|, & (3.24)\\ &\leq \lambda \|\tilde{g}(z)\| + (1-\lambda) \|\tilde{g}(y)\|. & (3.25) \end{aligned}$$

Note that $0 \leq \tilde{g}(\lambda z + (1-\lambda)y) \leq \lambda \tilde{g}(z) + (1-\lambda)\tilde{g}(y)$, where first inequality is because of the pointwise maximum in the definition of $\tilde{g}$ and second inequality due to convexity assumption of the function g. Therefore, implication (3.24) comes from the fact that $0 \leq z \leq y \implies \|z\| \leq \|y\|$. Implication (3.25) comes from Cauchy-Schwartz inequality. Hence clearly, l is convex if g and therefore, $\tilde{g}$ is convex. Therefore, we can reduce (3.2) to the following finite-dimensional convex optimisation problem

$$\inf_{x \in \mathbb{R}^d} \|\tilde{g}(x)\| \, \epsilon_N(\beta) + \frac{1}{N} \sum_{k=1}^{N} f(x, \widehat{\xi}_k). \tag{3.26}$$

Remark 3.1.8. *Examples of functions for which we can find a finite optimum value using the above reformulation are:*

(i) Any function affine in ξ.

(ii) Any continuous function $f : \mathcal{X} \times \Xi \mapsto \mathbb{R}$ if Ξ is bounded.

Note that both the equations (3.12) and (3.26) can be solved by each agent individually by using the data available to it. However, such a local solution may have an inferior out-of-sample guarantee and worse bias because all the data points, distributed all over the network, are not considered by any one agent while optimising (3.2). Therefore, in the next section, we frame and solve this DRO problem in distributed settings.

Chapter 4

Distributed algorithm analysis

In this chapter, in Section 4.1, we reformulate the central optimisation problem (3.12) to have a structure that confirms to the distributed optimisation framework explained in section 2.4. In Section 4.2, we construct an augmented convex-concave Lagrangian function of the reformulated problem, such that, its saddle point have one-to-one correspondence with the optimisers of the reformulated problem. In Section 4.3, using results from [6], we design the saddle-point dynamics for the augmented Lagrangian and prove asymptotic convergence of the trajectories to the saddle points of the Lagrangian and hence, to the optimisers of the original stochastic optimisation problem (3.12). At the end, we give the component-wise expressions for the distributed algorithm.

4.1 Reformulation as distributed optimisation problem

For the optimisation problem at hand, x is the global variable and our goal is that n agents are able to find the optimal x^* that minimises the optimisation problem in (3.12) while interacting over a constrained communication graph $\mathcal{G}$. We let each agent $i \in [n]$ maintains a copy of x, denoted by $x_i \in \mathbb{R}^d$, which is its decision variable for its local optimisation problem and ensure, using additional constraints, that after some iterations, all the $x_i's$ come to a consensus. This consensus value is then our minimiser x^*.
One possible reformulation of the central convex optimisation problem (3.12) to a convex optimization problem satisfying our distributed algorithm framework is:

$$\inf_{x_v} \quad c + \frac{1}{N}\sum_{k=1}^{N} f(x_{v_k}, \widehat{\xi}_k) \tag{4.1a}$$

$$\text{subject to} \quad (\mathbf{L} \otimes \mathbf{I}_d)x_v = \mathbf{0}_{nd}. \tag{4.1b}$$

where $x_v := (x_1; x_2; .....; x_n)$, $v_k \in [n]$ represents the agent holding the k-th sample $\widehat{\xi}_k$ of the dataset; $\mathbf{L} \in \mathbb{R}^{n\times n}$ is the Laplacian of the graph $\mathcal{G}$ and $c := \kappa\epsilon_N(\beta)$ which can be considered a constant dependent on the objective function, the desired reliability, and the metric used.
The constraint (4.1b) is included to ensure consensus among all the agents over the optimiser value ${x_v}^*$. We could impose any other constraint such as based on the *incidence matrix*, *edge Laplacian* or any constraint driving the individual $x_i's$ to their average value. In general, one prefers using a constraint that maintains sparsity in the problem structure.

Lemma 4.1.1 (One-to-one correspondence between optimizers of (3.6) and (4.1))**.** *The following holds:*

(i) If x^* is an optimizer of (3.12)*, then $(\mathbf{1}_n \otimes x^*)$ is an optimizer of* (4.1)*.*

(ii) If $x_v{}^*$ is an optimizer of (4.1), then there exists an optimiser x^* of (3.12) such that $x_v{}^* = \mathbf{1}_n \otimes x^*$.

The proof follows from the one given in [9]. Therefore, from Lemma 4.1.1, we get that the constraint (4.1b) ensures consensus and that (4.1) is equivalent to (3.12).
Note that the objective function (4.1a) can be written as

$$\inf_{x_v} F(x_v) := \inf_{x_v} \sum_{i=1}^{n} \Big(\frac{\kappa \epsilon_N(\beta)}{n} + \frac{1}{N} \sum_{k:\widehat{\xi}_k \in \widehat{\Xi}_i} f(x_i, \widehat{\xi}_k) \Big), \quad \forall i \in [n].$$

Therefore, the problem (4.1) has the adequate structure from a distributed optimization viewpoint,i.e., the objective function can be broken down to an aggregate of locally computable objective functions and constraints.

4.2 Lagrangian, Augmented Lagrangian and saddle points

In this section, we identify an appropriate augmented version of the Lagrangian function of (4.1) such that there is equivalence between its saddle points and the primal-dual optimizers of (4.1). The Lagrangian of (4.1) is $L : \mathbb{R}^{nd} \times \mathbb{R}^{nd} \longrightarrow \mathbb{R}$,

$$L(x_v, \eta) := F(x_v) + \eta^T (\mathbf{L} \otimes \mathbf{I}_d) x_v, \tag{4.2}$$

where $\eta \in \mathbb{R}^{nd}$ is the dual variable corresponding to the equality constraint (4.1b). L is convex-concave in (x_v, η).

Lemma 4.2.1. *(Min-max equality for L): The set of saddle points of L over the domain $\mathbb{R}^{nd} \times \mathbb{R}^{nd}$ is nonempty and*

$$\inf_{x_v} \sup_{\eta} L(x_v, \eta) = \sup_{\eta} \inf_{x_v} L(x_v, \eta) \tag{4.3}$$

Proof. The refined Slater conditions for strong duality in convex optimisation problem reduces to feasibility when the constraints are all linear equalities [4, Section 5.2.3]. In other words, for a convex optimisation problem, feasibility of primal problem implies strong duality. Also, primal problem is indeed feasible in our case since our objective function is proper. ■

Lemma 4.2.2. *(Strong duality for L): Lemma (4.2) implies the following:*
(i) If $(\overline{x}_v, \overline{\eta})$ is a saddle point of L over $\mathbb{R}^{nd} \times \mathbb{R}^{nd}$, then $\overline{x}_v$ is an optimizer of (4.1).
(ii) If $\overline{x}_v$ is an optimizer of (4.1), then there exists $\overline{\eta}$ such that $(\overline{x}_v, \overline{\eta})$ is a saddle point of L over $\mathbb{R}^{nd} \times \mathbb{R}^{nd}$.

Proof. We prove (i) by contradiction. Lets assume that $(\overline{x}_v, \overline{\eta}) \in \text{Saddle}(L)$ and $\overline{x}_v$ is not an optimiser of (4.1). Since we assume that there exists a finite optimiser of (3.12), let x^* be an optimiser of (3.12). Then by Lemma (4.1.1), we get that $F(\mathbf{1}_d \otimes x^*) < F(\overline{x}_v)$. We also know that $(\mathbf{L} \otimes \mathbf{1}_d)\overline{x}_v = \mathbf{0}_{nd}$, or else $\sup_\eta L(\overline{x}_v, \overline{\eta}) = +\infty$. Using the above two conclusions and putting them in (4.2), we get that $L(\mathbf{1}_d \otimes x^*, \overline{\eta}) < L(\overline{x}_v, \overline{\eta})$. This contradicts that $(\overline{x}_v, \overline{\eta}) \in \text{Saddle}(L)$. Hence, $(\overline{x}_v, \overline{\eta}) \in \text{Saddle}(L)$ implies that $(\overline{x}_v)$ is an optimiser of (4.1).
For (ii), lets assume that $\overline{x}_v$ is an optimiser of (4.1). Therefore, $F(\overline{x}_v) < F(x_v), \forall x_v \in \mathbb{R}^{nd}$. For an $\overline{\eta}$ such that $(\mathbf{L} \otimes \mathbf{1}_d)\overline{\eta} = \mathbf{0}_{nd}$, we get that $L(\overline{x}_v, \overline{\eta}) \leq L(x_v, \overline{\eta}), \forall x_v \in \mathbb{R}^{nd}$. Therefore, $(\overline{x}_v, \overline{\eta})$ is a saddle point of L. ■

Knowing that L admits saddle points corresponding to the optimisers of our optimisation problem, we can write the saddle-point dynamics for the Lagrangian L as a distributed algorithm and find the optimizers. We augment the Lagrangian with quadratic terms in the primal variables. Augmentation is helpful in giving steeper curvature to the Lagrangian, faster convergence rate and better behaved dynamics of the system in terms of less damping. One can augment the Lagrangian using any strictly convex term of the primal variable but typically one prefers to use a term that retains sparsity of the original Lagrangian as writing a distributed algorithm with a non-sparse augmentation may not be possible. Let the augmented Lagrangian $L_{aug} : (\mathbb{R}^{nd} \times \mathbb{R}^{nd}) \longrightarrow \overline{\mathbb{R}}$ be

$$L_{aug}(x_v, \eta) := L(x_v, \eta) + \frac{1}{2} {x_v}^T (\mathbf{L} \otimes \mathbf{I}_d) x_v. \tag{4.4}$$

The augmented Lagrangian is also convex-concave in (x_v, η) and admits saddle points corresponding to the optimisers of our optimisation problem. In addition, we have that the $\text{Saddle}(L) = \text{Saddle}(L_{aug})$, which is shown below.

Lemma 4.2.3. *(Saddle points of L and L_{aug} are the same): A point (x_v^*, η^*) is a saddle point of L over $(\mathbf{R}^{nd} \times \mathbf{R}^{nd})$ if and only if it is a saddle point of L_{aug} over $(\mathbf{R}^{nd} \times \mathbf{R}^{nd})$.*

Proof. Let $(x_v, \eta) \in \text{Saddle}(L)$, then since $x_v = (\mathbf{1}_n \otimes x)$, $\nabla_{x_v} L(x_v, \eta) = 0 \implies \nabla_{x_v} L_{aug}(x_v, \eta) = 0$. For the same reason, for $(x_v, \eta) \in \text{Saddle}(L_{aug})$, $(x_v, \eta) \in \text{Saddle}(L)$. ■

In addition, we have the following result which shows that for each $(x^*, z^*) \in \text{Saddle}(L_{aug})$, if $L_{aug}(x, z^*) = L_{aug}(x^*, z^*)$, then $(x, z^*) \in \text{Saddle}(L_{aug})$.

Lemma 4.2.4. *For L_{aug} defined in (4.4) consider any $({x_v}^*, \eta^*) \in \text{Saddle}(L_{aug})$. If $x_v \in \mathbb{R}^{nd}$ such that $L_{aug}(x_v, \eta^*) = L_{aug}({x_v}^*, \eta^*)$, then (x_v, η^*) is a saddle point of L_{aug}.*

Proof. The set of saddle points of L_{aug} can be characterised by using first-order conditions as

$$\left\{ (x_v, \eta) \in \mathbb{R}^{nd} \times \mathbb{R}^{nd} | x_v = \mathbf{1}_n \otimes x, x \in \mathbb{R}^d; \eta^T (\mathbf{L} \otimes \mathbf{I}_d) = -\nabla_{x_v} F(x_v) - (\mathbf{L} \otimes \mathbf{I}_d) x_v \right\}. \tag{4.5}$$

Substituting $x_v = \mathbf{1}_n \otimes x$ in the second constraint, we get

$$\eta^T (\mathbf{L} \otimes \mathbf{I}_d) = -\nabla_{x_v} F(\mathbf{1}_n \otimes x). \tag{4.6}$$

Consider $(x_v^*, \eta^*) \in \text{Saddle}(L_{aug})$ and any $x_v^{'} \in \mathbb{R}^{nd}$ such that $L_{aug}({x_v}^*, \eta^*) = L_{aug}({x_v}^{'}, \eta^*)$. This equality translates to

$$F(\mathbf{1}_n \otimes x^*) = F({x_v}^{'}) + {\eta^*}^T (\mathbf{L} \otimes \mathbf{I}_d) {x_v}^{'} + \frac{1}{2} {x_v}^{'T} (\mathbf{L} \otimes \mathbf{I}_d) {x_v}^{'} \tag{4.7}$$

$$\implies F(\mathbf{1}_n \otimes x^*) = F({x_v}^{'}) - \nabla_{x_v} F(\mathbf{1}_n \otimes x^*) {x_v}^{'} + \frac{1}{2} {x_v}^{'T} (\mathbf{L} \otimes \mathbf{I}_d) {x_v}^{'} \tag{4.8}$$

Implication (4.8) is obtained by substituting ${\eta^*}^T (\mathbf{L} \otimes \mathbf{I}_d) = -\nabla_{{x_v}^*} F(\mathbf{1}_n \otimes x^*)$ from (4.6). From the first-order convexity condition, we have

$$F({x_v}^{'}) \geq F({x_v}^*) + \nabla_{x_v} F({x_v}^*)^T ({x_v}^{'} - {x_v}^*) \tag{4.9}$$

$$= F(\mathbf{1}_n \otimes x^*) + \nabla_{x_v} F(\mathbf{1}_n \otimes x^*)^T ({x_v}^{'} - (\mathbf{1}_n \otimes x^*)) \tag{4.10}$$

Implication (4.10) is because $(x_v^*, \eta^*) \in \text{Saddle}(L_{aug}) \implies {x_v}^* = \mathbf{1}_n \otimes x^*$. Substituting $F(\mathbf{1}_n \otimes x^*)$ from (4.8) in the above inequality and canceling terms, we get

$$\frac{1}{2} {x_v}^{'T} (\mathbf{L} \otimes \mathbf{I}_d) {x_v}^{'} - \nabla_{x_v} F(\mathbf{1}_n \otimes x^*)^T (\mathbf{1}_n \otimes x^*) \leq \mathbf{0} \tag{4.11}$$

$$\implies \frac{1}{2}{x_v}'^T(\mathbf{L}\otimes\mathbf{I}_d)x_v{}' \leq \mathbf{0} \tag{4.12}$$

$$\implies x_v{}' = \mathbf{1}_n \otimes x' \tag{4.13}$$

Implication (4.12) comes from (4.8) and equality (4.13) because $(\mathbf{L}\otimes\mathbf{I}_d)$ is positive semi-definite. Therefore, the first constraint for x_v to belong in Saddle(L) is satisfied. For the second condition, we need to show that:

$$\eta^{*T}(\mathbf{L}\otimes\mathbf{I}_d) = -\nabla_{x_v}F(\mathbf{1}_n\otimes x') \implies (x_v', \eta^*) \in \text{Saddle}(L_{aug}). \tag{4.14}$$

We know that $x_v = (\mathbf{1}_n \otimes x')$ minimises $L_{aug}(x_v, \eta^*)$ i.e.$(\mathbf{1}_n \otimes x')$ is a critical point of L_{aug}, therefore $\nabla_{x_v}L_{aug}({x_v}', {\eta_v}^*) = 0$

$$\implies \nabla_{x_v}F(\mathbf{1}_n\otimes x')\eta^* + \eta^{*T}(\mathbf{L}\otimes\mathbf{I}_d) + (\mathbf{L}\otimes\mathbf{I}_d)(\mathbf{1}_n\otimes x') = 0 \tag{4.15}$$

$$\implies \nabla_{x_v}F(\mathbf{1}_n\otimes x')\eta^* = -(\mathbf{L}\otimes\mathbf{I}_d)\eta^* \tag{4.16}$$

Hence, $(x_v', \eta^*) \in \text{Saddle}(L_{aug})$. ■

Note that $L({x_v}^*, \eta^*) = L(x_v, \eta^*)$ does not necessarily imply $(x_v, \eta^*) \in \text{Saddle}(L)$, unless L has special characteristics such as strict convexity in x_v. Therefore, augmenting the Lagrangian was necessary.

4.3 System dynamics and convergence analysis

The saddle-point dynamics consists of gradient-descent of L_{aug} in the convex variables and gradient-ascent in the concave ones, i.e.,

$$\frac{dx_v}{dt} = -\nabla_{x_v}L_{aug}(x_v, \eta) \tag{4.17a}$$

$$\frac{d\eta}{dt} = \nabla_{\eta}L_{aug}(x_v, \eta) \tag{4.17b}$$

The next result guarantees the asymptotic stability of every $(x_v, \eta) \in \mathit{Saddle}(L_{aug})$ under the saddle-point dynamics X_{sp} (gradient-descent in the first variable x_v, and gradient-ascent in the second variable ξ).

Theorem 4.3.1. *(Global asymptotic stability of the set of saddle points via convexity-linearity): For the C^1 function L_{aug} defined in* (4.4)*,* Saddle(L_{aug}) *is globally asymptotically stable under the saddle-point dynamics X_{sp}* (4.17) *and convergence of trajectories is to a point.*

Proof. We know that L_{aug} is convex-concave in (x_v, η) and linear in η. Lemma 4.2.4 shows that L_{aug} satisfies that for each $(x^*, \eta^*) \in \text{Saddle}(L_{aug})$, if $L_{aug}(x, \eta^*) = L_{aug}(x^*, \eta^*)$, then $(x, \eta^*) \in \text{Saddle}(L_{aug})$. Therefore, from [6, Corollary 4.5], we conclude that asymptotic convergence of the trajectories of L_{aug} under the saddle-point dynamics X_{sp} to Saddle(L_{aug}) is guaranteed. Nonetheless, we present a concise proof based on LaSalle Invariance Principle for the sake of completeness.
Let (x_v^*, η^*) be the equilibrium point of L_{aug} then $(\mathbf{L}\otimes\mathbf{I}_d)x_v^* = \mathbf{0}_{nd}$ and $(x_v^*, \eta^*) \in \text{Saddle}(L)$. Consider the function $V : \mathbf{R}^{nd}\times\mathbf{R}^{nd} \mapsto \mathbf{R}_{\geq 0}$,

$$V(x_v, \eta) := \frac{1}{2}(\|x_v - x_v^*\|^2 + \|\eta - \eta^*\|^2).$$

The Lie derivative of V along the dynamics for some $(x, z) \in \mathcal{N}$, where $\mathcal{N}$ is some neighbourhood of (x^*, z^*), is:

$$\mathcal{L}_{X_{SP}}V(x_v, \eta) = -(x_v - {x_v}^*)^T\nabla_{x_v}L_{aug}(x_v, \eta) + (\eta - \eta^*)^T\nabla_{\eta}L_{aug}(x_v, \eta) \tag{4.18a}$$

$$\leq L_{aug}(x_v{}^*, \eta) - L_{aug}(x_v, \eta) + L_{aug}(x_v, \eta) + L_{aug}(x_v, \eta^*) \tag{4.18b}$$
$$= L_{aug}(x_v{}^*, \eta) - L_{aug}(x_v{}^*, \eta^*) + L_{aug}(x_v{}^*, \eta^*) + L_{aug}(x_v, \eta^*) \leq 0 \tag{4.18c}$$

Implication (4.18b) comes by applying first order convexity-concavity condition on L_{aug} and (4.18c), from the fact that $(x_v{}^*, \eta^*)$ is a saddle point of L_{aug}. From (4.18c), we can conclude that for a sufficiently small $\alpha > 0$, such that the alpha sublevel set $V^{-1}(\leq \alpha) \subset \mathcal{N}$, $V^{-1}(\leq \alpha)$ is compact and positive invariant under X_{SP}. Therefore, from *LaSalle Invariance Principle* (2.6.1), any trajectory starting in $V^{-1}(\leq \alpha)$ converges to the largest invariant set, M, where $M := \{(x_v, z) \in V^{-}(\leq \alpha) | \mathcal{L}_{X_{SP}} V(x_v, \eta) = 0\}$.
Now, for a $(x, \eta) \in M$, using (4.18c), we get

$$\mathcal{L}_{X_{SP}} V(x_v, \eta) = 0 \implies L_{aug}(x_v{}^*, \eta) = L_{aug}(x_v{}^*, \eta^*) = L_{aug}(x_v, \eta^*).$$

Therefore, Lemma 4.2.4 in 4.3.1 gives that $(x_v, \eta^*) \in \text{Saddle}(L_{aug})$ and convexity-linearity of L_{aug} gives that $\nabla_\eta L_{aug}(x_v, \eta) = \nabla_\eta L_{aug}(x_v, \eta^*) = 0, \forall (x_v, \eta) \in M$ since L is linear in η. This implies that $\eta(t) = \eta$ for all $t \in [0, +\infty)$ and $\dot{x}_v(t) = -\nabla_{x_v} L_{aug}(x_v(t), \eta)$ corresponds to the gradient descent of just a convex function. Clearly, $x_v(t)$ converges to the minimiser $x_v{}'$ such that $\nabla_{x_v} L_{aug}(x_v{}', \eta) = 0$.
Since, $\nabla_{x_v} L_{aug}(x_v{}', \eta) = 0$ and $\nabla_\eta L_{aug}(x_v{}', \eta) = 0$, we get that $(x_v{}', \eta) \in \text{Saddle}(L_{aug})$. From [6], Lemma A.1, we know that $L_{aug|\,\text{Saddle}(L_{aug})}$ is constant which implies that $L_{aug}(x_v{}', \eta) = L_{aug}(x_v{}^*, \eta^*)$.
From the chain of above arguments, we get that $L_{aug}(x_v(t), \eta) \to L_{aug}(x_v{}^*, \eta^*)$.
In addition, from (4.18a), we get:

$$\mathcal{L}_{X_{SP}} V(x_v(t), \eta) = 0 = -(x_v - x_v{}^*)^T \nabla_{x_v} L_{aug}(x_v(t), \eta) \leq L_{aug}(x_v{}^*, \eta) - L_{aug}(x_v(t), \eta)$$

$$\implies L_{aug}(x_v(t), \eta) \leq L_{aug}(x_v{}^*, \eta) = L_{aug}(x_v{}^*, \eta^*), \forall t \in [0, +\infty).$$

So, we know that the function $L_{aug}(x_v(t), \eta)$ following gradient descent in x converges to $L(x_v{}^*, \eta^*)$ which is the maximum of all $L_{aug}(x_v, \eta)$, for all $(x_v, \eta) \in M$. This is possible only if $L_{aug}(x_v(t), \eta)$ is equal to $L_{aug}(x_v{}^*, \eta^*)$ and hence a constant for all $t \in [0, +\infty)$. Therefore, $\nabla_{x_v} L_{aug}(x_v(t), \eta) = 0, \forall t \in [0, +\infty)$.
Hence, $(x_v, \eta) \in \text{Saddle}(L_{aug}), \forall (x_v, \eta) \in M$ and $\text{Saddle}(L_{aug})$ is asymptotically stable. The characteristics of $\text{Saddle}(L)$ are given in (4.5) ■

4.3.1 Distributed implementation of saddle-point dynamics

Let the components of the dual variable η be denoted as $\eta = (\eta_1; \eta_2; \ldots; \eta_n)$ such that each agent $i \in [n]$ maintains $\eta_i \in \mathbb{R}^d$. For implementing X_{sp}, each agent i maintains and updates the variables (x_i, η_i). The collection of these variables for all $i \in [n]$ forms (x_v, η). From (4.17), the dynamics of variables maintained by i is

$$\frac{dx_i}{dt} = -\frac{1}{N} \sum_{k \in \mathcal{K}_i} \nabla_x f(x_i, \widehat{\xi}_k) - \sum_{j \in \mathcal{N}_i} a_{ij} \big((\eta_i - \eta_j) + (x_i - x_j) \big), \tag{4.19}$$

$$\frac{d\eta_i}{dt} = \sum_{j \in \mathcal{N}_i} a_{ij} (x_i - x_j). \tag{4.20}$$

Remark 4.3.2. *The dynamics for reformulation* (3.26) *will be*

$$\frac{dx_i}{dt} = -\frac{1}{N} \sum_{k \in \mathcal{K}_i} \nabla_x \big(\|\tilde{g}(x)\| + f(x_i, \widehat{\xi}_k) \big) - \sum_{j \in \mathcal{N}_i} a_{ij} \big((\eta_i - \eta_j) + (x_i - x_j) \big), \tag{4.21}$$

$$\frac{d\eta_i}{dt} = \sum_{j \in \mathcal{N}_i} a_{ij}(x_i - x_j) \tag{4.22}$$

where $\tilde{g}(x)_i := \max\{\big(\nabla_\xi f(x, \xi_k)\big)_i, 0\}$, $i \in [n]$.

In the next chapter, we show the simulation results for the evolution of the primal variable and the objective function values obtained as per the above dynamics for various convex objective functions and analyse these results.

Chapter 5

Simulation results

In this section, we verify the distributed algorithm (4.17) to find a data-driven solution for various objective functions in each subsection. The network configuration for each objective function is the same and as follows:
We take $n = 10$ agents communicating over a graph with the structure of an undirected ring with additional edges $\{(1,4),(2,5),(3,7),(6,10)\}$ with binary weights.

Least squares problem

We consider a linear regression problem with quadratic loss function. We use finite data points to find an affine predictor $x \in \mathbb{R}^5$ such that, for a new unseen sample $\xi = (w, y)$, the predictor $x^\top(w;1)$ is equal to y. A possible strategy for this, is to solve the following

$$\inf_x \mathbb{E}_{\mathbb{P}}\Big[f(x,w,y)\Big] \tag{5.1}$$

where $\mathbb{P}$ is the probability distribution of the data (w, y) and $f : \mathbb{R}^5 \times \mathbb{R}^4 \times \mathbb{R} \to \mathbb{R}$ is the quadratic loss function, i.e., $f(x,w,y) = (x^\top(w;1) - y)^2$ which is very commonly used for regression problems. The quadratic loss function is smooth, convex and penalises the difference between the prediction $x^\top(w;1)$ and the actual label y. We collect data as follows:
Each agent collects 30 i.i.d data points of the form $\widehat{\xi}^k = (\widehat{w}^k, \widehat{y}^k)$ where $\widehat{w}^k \in \mathbb{R}^4$ is the input and $\widehat{y}^k \in \mathbb{R}$ the output label. Hence all agents together collect $N = 300$ total number of samples. The input vector w is assumed to have a standard multivariate normal distribution, that is, zero mean and covariance as the identity matrix I_4. The output y is assigned values $y = [1,4,3,2] * w + v$ where v is a random variable, uniformly distributed over the interval $[-1,1]$. This defines completely the distribution $\mathbb{P}$ of (w, y). We choose the significance value, $\beta = 0.95 \in (0,1)$. Therefore, calculating corresponding ϵ for $N = 300$ according to (2.9) we get $\epsilon_N(\beta) = 0.05$ approximately. Note that we cannot exactly calculate $\epsilon_N(\beta)$ for a β unless we know the parameters a, b and m for the distribution $\mathbb{P}$ 2.8. We make this value β known to all the agents, so that they can compute $\epsilon_N(\beta)$ beforehand.
The evolution of x_v, for a suitable step size, is shown in Figure 5.

Convex function affine in uncertainty

We consider a minimisation problem of following convex functions

(i) $f_1 : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ defined as $f_1(x,\xi) := \xi x^2$,

(ii) $f_2 : \mathbb{R}^2 \times \mathbb{R} \mapsto \mathbb{R}$ defined as $f_2(x,\xi) := \xi(x^T P x + bx)$, where $P = [0.75, 0; 00.25] \in \mathbb{R}^{2\times 2}$ is a positive semidefinite matrix and $b = [0.250.75] \in \mathbb{R}^2$.

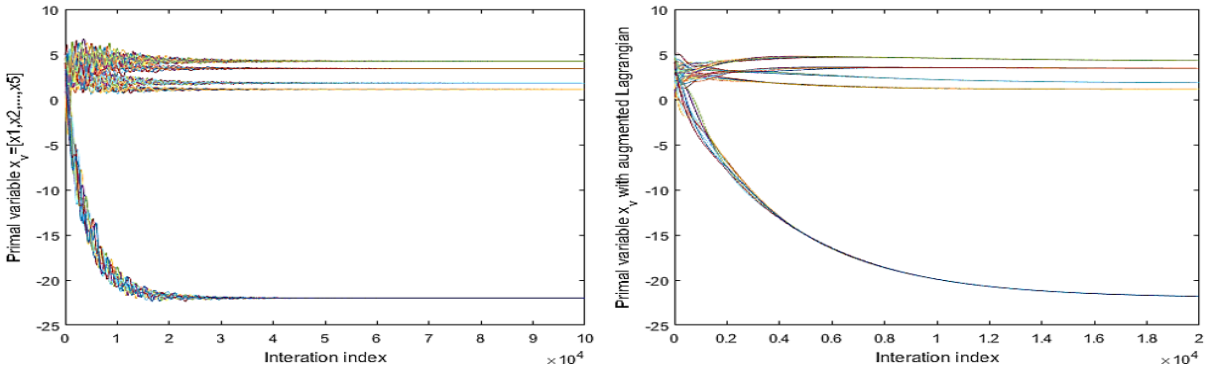

Figure 5.1: Evolution of the dynamics (4.17) to solve regression problem (5.1) in a data driven manner.
Plot 5(a) and plot 5(b) depict the evolution of the primal variable of the distributed optimization problem (4.1) defined for objective function (5.1) with and without augmentation of the Lagrangian respectively. The initial condition $(x_{\mathrm{v}}(0), \lambda_{\mathrm{v}}(0))$ is chosen randomly from the set $[0,5]^{50} \times [30,80]^{10}$ and $\eta(0) = \mathbf{0}_{50}$. The primal variable converge to $x_{\mathrm{v}} = \mathbf{1}_{10} \otimes x^*$ with $x^* = [1.1284, 1.8253, 3.4423, 4.2458, -21.9913]$. This is an equilibrium point of X_{sp} and an optimizer of (4.1). One can see that the agents do not reach consensus early in the execution when the Lagrangian is not augmented and we observe a lot of oscillation in the curves of x_{v} in plot 5(a). While in the second 5(b), the agents converge to a consensus value smoothly and much earlier.

The ξ in both cases has a standard normal distribution, that is, zero mean and unit covariance. We use N data points (300 in total with each agent gathering 30 of them) to find the minimiser x^* of f such that it satisfies finite sample guarantee, i.e., for a new unseen sample $\xi^{'}$ which we take from a validation dataset kept separately from the training dataset, $f(x^*, \xi^{'})$ is less than $\widehat{J}_N := \inf_{x \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathcal{B}_{\epsilon_N(\beta)}} \mathbb{E}^{\mathbb{Q}}[f(x, \xi)]$ with probability greater than $1 - \beta$. For this, we minimise the function using both reformulations derived in Chapter 3 and compare the results. We keep $\epsilon_N(\beta) = 0.05$ as for the previous problem.
The evolution of x_v, for a constant step size as per both reformulation (3.12) and reformulation (3.26), is shown in Figure 5 plot.

Convex function affine in uncertainty with streaming data accumulation

We consider a minimisation problem of convex function f_2 (ii) from the previous subsection under the following data accumulation setup:

(i) each of the n agents initially have a local dataset of 30 samples. N_{stream} total number of more data points, one at a time, are given to any one of the n agents by random selection. New data points are adding at randomly chosen time instants.

To find the minimiser x^* of f such that it satisfies finite sample guarantee, we minimise the function using both the reformulations derived in Chapter 3 and compare the results. We keep $\epsilon_N(\beta) = 0.05$ as in the previous problem.
The evolution of x_v, for a constant step size as per both reformulation (3.12) and reformulation (3.26), is shown in Figure 5 plots.

Simulations on similar connected graphs (undirected and directed, weight-balanced graphs) and non-differentiable convex functions were also carried out to find similar results as those that have been shown in the coming sections. In case of streaming data, simulations were done for different rates and amounts of data addition and dynamics were still found to converge.

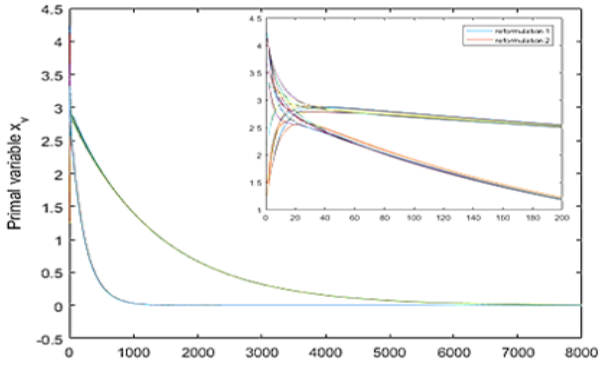

Figure 5.2: Evolution of the dynamics (4.17) to minimise function f_1 (defined in (i)) in a data driven manner.
Plot above depicts the evolution of the scalar primal variable of the distributed optimization problem (4.1) defined for objective function $f_1(x, \xi) := \xi x^2$ with reformulations (3.12) and (3.26). A smaller window inside the plot shows the zoomed in version of trajectories of x_v components, so that we can view the dynamics better. The gradient descent and ascent dynamics on the reformulation (3.26) are smoother and converge much faster as compared to the same dynamics on reformulation (3.12).

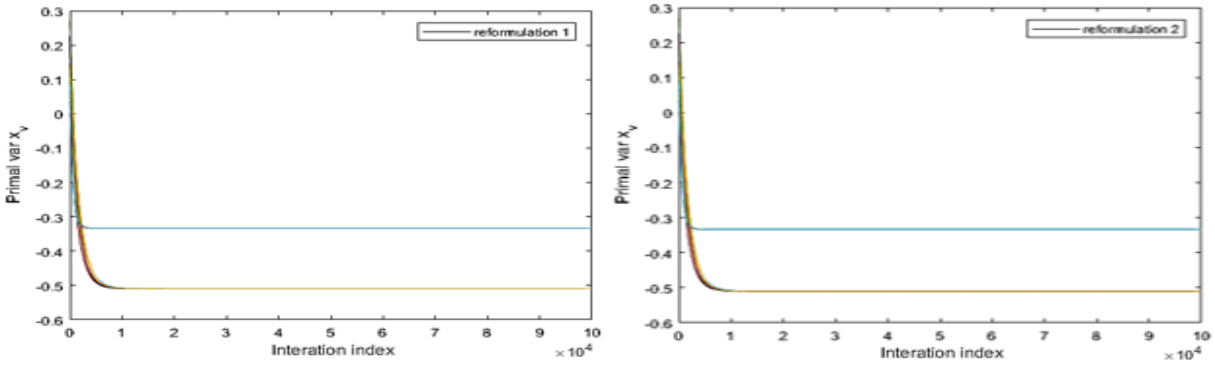

Figure 5.3: Evolution of the dynamics (4.17) to minimise function f_2 (defined in (ii)) in a data driven manner.
Plots above depict the evolution of the scalar primal variable of the distributed optimization problem (4.1) defined for objective function $f_2(x, \xi) := \xi(x^T P x + bx)$ with reformulations (3.12) and (3.26). The gradient descent and ascent dynamics on the reformulation (3.26) converge to give smaller certificate $\widehat{J}_N = -0.006$ as compared to the same dynamics on reformulation (3.12) which gives 0.117 as certificate.

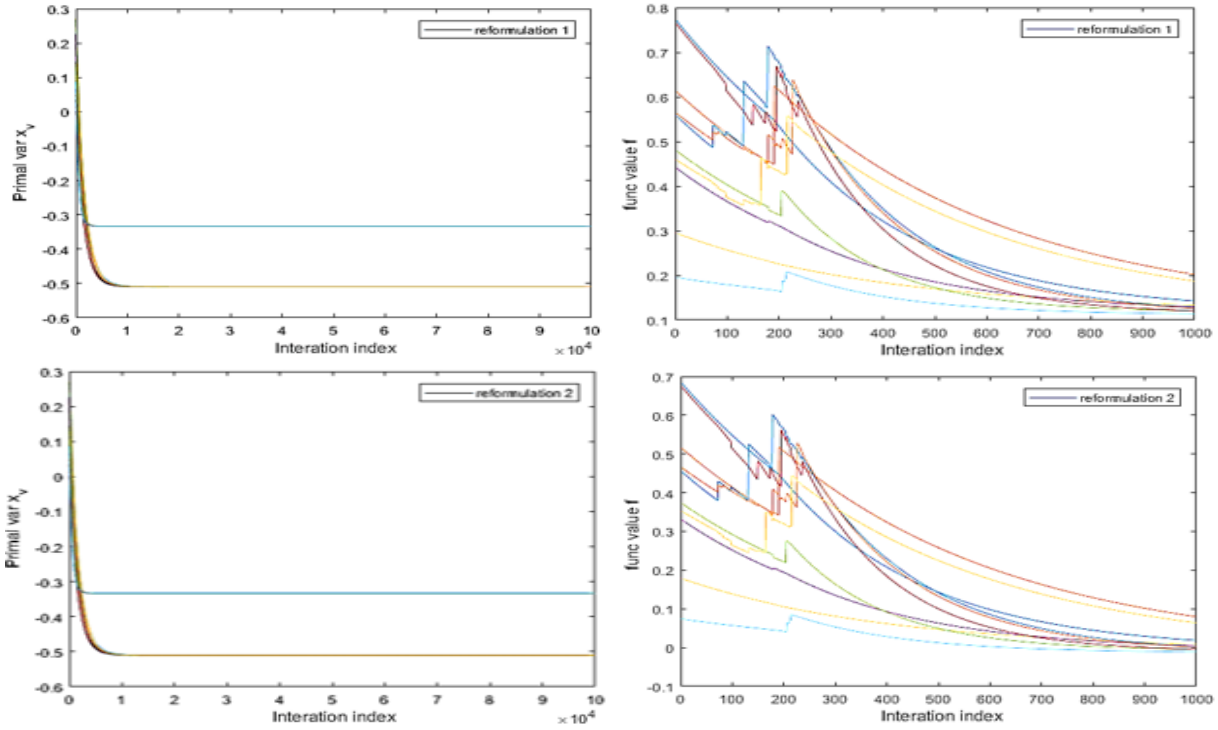

Figure 5.4: The left plots (a) show the evolution of the dynamics (4.17) to minimise function f_2 (ii) in a data driven manner with new data points being randomly added online. The right plots (b) show the value of f_2 with the current x_v and the data points representing the uncertainty ξ. One can see the jumps in the value of objective function as a new data point is added. The dynamics still converge in almost as much time as without new data points addition. The plot is shown till the first 1000 iterations to make the jumps in the objective function value visible. Plots 5(a) depict the evolution of the scalar primal variable of the distributed optimization problem (4.1) defined for objective function $f_2(x, \xi) := \xi(x^T P x + bx)$ with reformulations (3.6) and (3.26). The saddle point dynamics on the both the reformulations (3.6) and (3.26) converge to almost the same minimisers as in the previous minimisation with no new streamed data inclusion. Hence, we verify that the proposed algorithm converges to the minimisers of the optimisation problem even under streaming data. The new data points were added in the initial few iterations to be able to perturb the dynamics while it is converging, adding them later does not disturb the dynamics at all.

Chapter 6

Conclusions

In this thesis, we solve stochastic optimization problems cooperatively in multi-agent setting. We use DRO approach where we characterise the uncertainity set of the DRO using Wasserstein metric. To summarize, we do the following:

(i) reformulate the intractable DRO problem for a convex objective function into a convex problem, which can be solved in polynomial time and gives intuitive understanding.

(ii) design a distributed algorithm for this convex problem such that agents arrive at a common optimal solution by sharing without much network overhead.

(iii) formally establish asymptotic convergence of the dynamics derived by the distributed algorithm to the minimisers of the original stochastic optimization problem.

(iv) verify our analytical results and proposed algorithm by running simulations for some standard convex objective functions, also under streaming data settings.

Our discussion is largely useful for the class of objective functions which are affine in the uncertainty variable, which constitute a large chunk of functions used in finance for risk analysis. It also gives useful insight for much larger class of functions since piecewise affine functions frequently serve as approximations for smooth convex or concave objective functions.

6.1 Future Work

Future work will include theoretical bounds on the saddle point dynamics convergence rate under different data accumulation settings such as when new data becomes available in an online fashion. In this work, we only had uncertainity affecting the objective function, we would also like to analysis optimisation problems with network chance constraints and perturbations.

Bibliography

[1] S. Arora and B. Barak. Computational Complexity: A Modern Approach. Princeton University, 2007.

[2] K. J. Arrow, S. Karlin, and H. E. Scarf. A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, pages 201–209, 1958.

[3] D. Bertsimas, V. Gupta, and N. Kallus. Robust Sample Average Approximation. *Optimization and Control*, 2016.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] A. Cherukuri and J. Cortés. Cooperative data-driven distributionally robust optimization. *Optimization and Control*, pages 1368–1377, 2017.

[6] A. Cherukuri, B. Gharesifard, and J. Cortés. Saddle-point dynamics: conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, 55(1):486–511, 2017.

[7] M. Dyer and L. Stougie. Computational complexity of stochastic programming problems. *Mathematical Programming*, 2006.

[8] P. M. Esfahani and D. Kuhn. Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations. *Mathematical Programming*, pages 169–172, 2017.

[9] B. Gharesifard and J. Cortés. Distributed Continuous-Time Convex Optimization on Weight-Balanced Digraphs. *IEEE transactions on automatic control*, 59(3):781–786, 2014.

[10] G. A. Hanasusanto, D. Kuhn, and W. Wiesemann. A Comment on Computational Complexity of Stochastic Programming Problems. *Mathematical Programming*, pages 557–569, 2006.

[11] H. K. Khalil. *Nonlinear Systems*. Prentice Hall, 3 ed., 2002.

[12] K. Natarajan, M. Sim, and J. Uichanco. Asymmetry and Ambiguity in Newsvendor Models. *Optimisation Online*, 2008.

[13] L. N. Polyakova. On minimizing the sum of a convex function and a concave function,. Demyanov V.F. and Dixon L.C.W. (eds) Quasidifferential Calculus, Mathematical Programming Studies, 1986.

[14] A. Shapiro. On duality theory of conic linear problems, in Semi-Infinite Programming. *Optimization*, 55(6):535–543.

[15] A. Shapiro. Distributionlly Robust Stochatic Programming . Optimization Online, 2015.

[16] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014.

[17] D. Wozabal. Robustifying convex risk measures for linear portfolios: A nonparametric approach. *Operations Research*, 62(6):1302–1315, 2014.