Title
Stochastic Optimization: Complexity-Based Analysis and Development Engineering Applications

Author
Bugg, Caleb Xavier

Publication Date
2022

Peer reviewed|Thesis/dissertation

Stochastic Optimization: Complexity-Based Analysis and Development Engineering Applications

by

Caleb Xavier Bugg

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering- Industrial Engineering and Operations Research

and the Designated Emphasis

in

Development Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Anil Aswani, Chair
Professor Alper Atamturk
Professor Ilan Adler
Professor S. Shankar Sastry

Spring 2022

Stochastic Optimization: Complexity-Based Analysis and Development Engineering Applications

Abstract

Stochastic Optimization: Complexity-Based Analysis and Development Engineering Applications

by

Caleb Xavier Bugg

Doctor of Philosophy in Engineering- Industrial Engineering and Operations Research

and the Designated Emphasis in

Development Engineering

University of California, Berkeley

Associate Professor Anil Aswani, Chair

This dissertation work combines two lines of work related to stochastic optimization, one focused on theoretical contributions to the field, and the other focused on modelling and philosophical contributions to the field. In the first line of work, we contribute to the theoretical underpinnings of two models that draw from data analysis, mathematical and statistical modeling, and optimization. In particular, there are several classes of machine learning problems which interface with stochastics and optimization, and, with the use of stochastic process theory and associated notions of complexity, we can recharacterize some of the theoretical expectations of these problems, which importantly model uncertainty. The improved results are gained from problem reformulations, improved analytical techniques, and a concerted effort to use the most recent advances in complexity analysis and their associated generalizations, and has implications for the construction and analysis of *societal scale models*. In the second line of work, we demonstrate why the techniques in stochastic optimization help form a mathematical and statistical basis that can solve large-scale societal problems over time, with precise engineering intervention focus. The main model researched is the Principal Agent Model (PAM). Considering sectors of the *social economy* as the principal and communities of people as the agents, we utilize development engineering and the associated economics of global poverty reduction to unveil the concept of *Optimal Intervention Theory* as a means to explore, adapt, and activate tractable solutions in specially-constrained environments. By tractable here, we mean interventions that last long beyond the intervention period, according to the education-, resource-, access-, and network-bases interventions afforded the impacted populations, especially those concerned with international development and global poverty alleviation.

To the Youths of this Millennium, whose seriousness and ardour for Mankind have led us to the precipice of the Dream. One Nation, Under God, with liberty and justice for all.

Contents

List of Figures

List of Tables

Acknowledgments

Thank you to my supervisor, Dr. Anil Aswani, for providing guidance and feedback throughout the dissertation. Thanks also to my dissertation committee for their feedback and guidance, the IEOR department, for their financial and academic support of my graduate tenure, and to the various colleagues across the mathematical sciences who have contributed to a rich intellectual graduate tenure. Lastly, I would like to the the Bugg clan, my wonderful family, and my Morehouse Math community, who have supported me every step of the way.

For contributions to the Tensor Completion work, I would like to thank Dr. Chen Chen, who is a faculty member of The Ohio State University's Department of Integrated Systems Engineering. For contributions to the Sample Average Approximation work, I would like to thank Dr. Deepak Rajan, who is a faculty member at UC Berkeley's Industrial Engineering and Operations Research (IEOR) department.

Chapter 1

Introduction

Our global community has found itself at an interesting time nexus, in which there are great changes occurring in both natural and human systems. Changes to human systems are brought about by the globalization of world populations through technology, adding pressure to our ability to efficiently allocate global resources. The ethical ideals of altruism and utilitarianism are paramount here, as a correct (engineering) response to this current amalgamated crisis requires selflessness and a desire to (1) do the most good for the greatest number of people on this planet, and (2) help and prioritize those who need the most help.

The prospect of repealing our current systems is an easier task than *replacing those systems with systems based in Justice, Equity, Diversity, and Inclusion (JEDI)* principles. To the task of replacing these systems, this thesis represents initial steps towards combining social, political, and economic (SPE) thinking with rigorous *stochastic optimization* theory, which is the basis of academic "buzzwords" like artificial intelligence, machine learning, and data science. In particular, following the title of Holland's lead of viewing the World's economy as an adaptive process [43], we posit a mathematical view of human and social systems that can, over time, *learn* its tasks better, and therefore improve *ad infinitum* to higher social and physical utility of economic resources.

The field of *development engineering* (DE) is concerned with many of these issues; however, there is a lack of focus on incentive theory, contract design, and contract enforcement that will be critical to the future success of DE. Operations Researchers and Systems Engineers are equipped with the mathematical and statistical tools necessary to present and collate methods from statistics and optimization, and our associated knowledge of their in-use efficiency, to begin the rigorous study of DE. Whereas DE contributes to literature and practice of Global Poverty Alleviation and International Development (GPA & ID) by considering the combination of economics and engineering science, OIT utilizes the theory and practice of additional social sciences, namely behavioral psychology, to capture social and political factors that impact human behavior and the adoption and use of goods and services beneficial to society.

This dissertation falls under the guise of Public Operations Research (Public OR), and the field of Machine Learning and AI for Social Good (ML4SG/AI4SG), which are closely connected to diversity, equity, and inclusion outcomes for social systems such as schools, healthcare systems, and housing and food security. Many of the practitioners of these fields and subfields focus on the two problems of (1) translating abstract, qualitative fairness and

equity goals into "math problems", and (2) extending the prediction and classification power of Machine Learning (ML) algorithms and techniques to intervention insights. Both of these things are likely impossible to do in a completely technical format; and thus will have to lend credence to human actors and decision makers through the formation of multidisciplinary research and implementation teams.

1.1 Overview of Thesis

The creation and use of Optimal Intervention Theory requires both knowledge of stochastic optimization, and its associated theory, models, and analysis techniques, as well as general system modelling knowledge, for which operations researchers are rigorously trained. In particular, stochastic optimization gives us a framework and theory to pose models for OIT, and a setting for engineering techniques and analysis to be applied to GPA & ID. Stochasticity affects both the analysis of our models and the philosophy of observing and responding to random, human behavior. In this way, this dissertation work combines two lines of work related to stochastic optimization, one focused on theoretical contributions to the field, and the other focused on modelling and philosophical contributions to the field.

The two stochastic optimization methods presented in this work are Nonnegative Tensor Completion (NNTC) and Sample Average Approximation (SAA). The next two chapters study these stochastic optimization methods. The tensor completion problem represents the work we can do with data structures which hold our realized random variables' value over time, again helping us to predict the future value random variables through completion, and also gives us insights through the direct use of factorization techniques, similar to singular value decomposition in the matrix world. The SAA chapter gives results for stochastic problems involving ℓ_1 constraints, and will be a direct part of the OIT model, and our sample bounds in that chapter will be relevant for estimating the principal's investments over the intervention period, which is discussed in the final chapter of this dissertation.

The fourth and fifth chapters present the philosophical and mathematical basis for OIT, respectively. The fourth chapter focuses on modelling, and gives our unified systems-based theory for Global Poverty Alleviation and International Development (GPA & ID), while the fifth chapter presents math models that are tailored to form our Quadratic Knapsack Problem, which is based on a dynamic version of the Principal Agent Model. The final chapter previews the impact of OIT through cost savings and shared benefits models, and concludes with ethical and pedagogical contributions to the field of development engineering and engineering ethics.

Tensor Completion

Unlike matrix completion, tensor completion does not have an algorithm that is known to achieve the information-theoretic sample complexity rate. Chapter 2 develops a new algorithm for the special case of completion for nonnegative tensors. We prove that our algorithm converges in a linear (in numerical tolerance) number of oracle steps, while achieving the information-theoretic rate. Our approach is to define a new norm for nonnegative tensors using the gauge of a particular 0-1 polytope; integer linear programming can, in turn, be used

to solve linear separation problems over this polytope. We combine this insight with a variant of the Frank-Wolfe algorithm to construct our numerical algorithm, and we demonstrate its effectiveness and scalability through computational experiments using a laptop on tensors with up to one-hundred million entries. This chapter is based on our work from [21].

Sample Average Approximation

Sample Average Approximation (SAA) is used to approximately solve stochastic optimization problems. In practice, SAA requires much fewer samples than predicted by existing theoretical bounds that ensure the SAA solution is close to optimal. Here, we derive new sample-size bounds for SAA that, for certain problems, are logarithmic (existing bounds are polynomial) in problem dimension. Notably, our new bounds provide a theoretical explanation for the success of SAA for many capacity- or budget-constrained optimization problems. This chapter based upon work from [20].

OIT

In chapters four and five, we present the basis and main model of OIT, an intervention model whose output is a dynamic optimal contract of interventions between a principal and agent under a time-indexed version of the quartic knapsack problem. The main thrust of adaptive processes is that decision makers (or processes) increase their knowledge by the cumulative experience of "doing while learning". The philosophical framework for optimal intervention theory is based in economic theory, with a special attention paid to behavioral psychology and other fields that contribute to the study of optimal incentive and contract design. OIT is the study of population *adoption* of *socially beneficial goods and services (SBGs).*

Chapter 2

Nonnegative Tensor Completion

Tensors generalize matrices. A tensor ψ of *order* p is $\psi \in \mathbb{R}^{r_1 \times \cdots \times r_p}$, where r_i is the *dimension* of the tensor in the i-th index, for $i = 1, \ldots, p$. Though related, many problems that are polynomial-time solvable for matrices are NP-hard for tensors. For instance, it is NP-hard to compute the rank of a tensor [41]. Tensor versions of the spectral norm, nuclear norm, and matrix singular value decomposition are also NP-hard to compute [41, 33].

2.1 Past Approaches to Tensor Completion

Tensor completion is the problem of observing (possibly with noise) a subset of entries of a tensor and then estimating the remaining entries based on an assumption of low-rankness. The tensor completion problem is encountered in a number of important applications, including computer vision [64, 112], regression with only categorical variables [6], healthcare [35, 27], and many other application domains [97].

Although the special case of matrix completion is now well understood, the tensor version of this problem has an unresolved tension. To date, no tensor completion algorithm has been shown to achieve the information-theoretic sample complexity rate. Namely, for a tensor completion problem on a rank k tensor with sample size n, the information theoretic rate for estimation error is $\sqrt{k \cdot \sum_i r_i / n}$ [35]. In fact, evidence suggests a computational barrier in which no polynomial-time algorithm can achieve this rate [9]. One set of approaches has polynomial-time computation but requires exponentially more samples than the information-theoretic rate [35, 71, 9, 70], whereas another set of approaches achieves the information-theoretic rate but (in order to attain global minima) requires solving NP-hard problems that lack numerical algorithms [23, 111, 110, 85].

However, algorithms that achieve the information-theoretic rate have been developed for a few special cases of tensors. Completion of nonnegative rank-1 tensors can be written as a convex optimization problem [6]. For symmetric orthogonal tensors, a variant of the Frank-Wolfe algorithm has been proposed [84]. Though this latter paper does not prove their algorithm achieves the information-theoretic rate, this can be shown using standard techniques [35]. This latter paper is closely related to the approach we take in this chapter, with one of the key differences being the design of a different separation oracle in order to support a different class of tensors.

2.2 Completion using Integer Programming

This chapter proposes a numerical algorithm for completion of nonnegative tensors that provably converges to a global minimum in a linear (in numerical tolerance) number of oracle steps, while achieving the information-theoretic rate. Nonnegative tensors are encountered in many applications. For instance, image and video data usually consists of nonnegative tensors [64, 112]. Notably, the image demosaicing problem that must be solved by nearly every digital camera [61] is an instance of a nonnegative tensor completion problem, though it has not previously been interpreted as such. Nonnegative tensor completion is also encountered in specific instances of recommender systems [97], healthcare applications [35, 27], and statistical regression contexts [6].

Our approach is to define a new norm for nonnegative tensors using the gauge of a specific 0-1 polytope that we construct. We prove that this norm acts as a convex surrogate for rank and has low statistical complexity (as measured by the Rademacher average for the tensor, viewed as a function from a set of indices to the corresponding entry), but is NP-hard to approximate to an arbitrary accuracy. Importantly, we prove that the tensor completion problem using this norm is able to achieve the information-theoretic rate in terms of sample complexity. However, the resulting tensor completion problem is NP-hard to solve. Nonetheless, because our new norm is defined using a 0-1 polytope, this means we can use integer linear optimization as a practical means to solve linear separation problems over the polytope. We combine this insight with a variant of the Frank-Wolfe algorithm to construct our numerical algorithm, and we demonstrate its effectiveness and scalability through numerical experiments.

2.3 Norm for Nonnegative Tensors

Consider a tensor $\psi \in \mathbb{R}^{r_1 \times \cdots \times r_p}$. To refer to a specific entry in the tensor we use the notation $\psi_x := \psi_{x_1,\ldots,x_p}$, where $x = (x_1, \ldots, x_p)$, and $x_i \in [r_i]$ denotes the value of the i-th index, with $[s] := \{1, \ldots, s\}$. Let $\rho = \sum_i r_i$, $\pi = \prod_i r_i$, and $\mathcal{R} = [r_1] \times \cdots \times [r_p]$.

A nonnegative rank-1 tensor is $\psi_x = \prod_{k=1}^{p} \theta^{(k)}_{x_k}$, where $\theta^{(k)} \in \mathbb{R}^{r_k}_+$ are nonnegative vectors indexed by the different values of $x_k \in [r_k]$. To simplify notation, we drop the superscript in $\theta^{(k)}_{x_k}$ and write this as θ_{x_k} when clear from the context.

For a nonnegative tensor ψ, its nonnegative rank is

$$\operatorname{rank}_+(\psi) = \min\{q \mid \psi = \textstyle\sum_{k=1}^{q} \psi^k, \psi^k \in \mathcal{B}_\infty \text{ for } k \in [q]\}, \tag{2.1}$$

where we define the ball of nonnegative rank-1 tensors whose maximum entry is $\lambda \in \mathbb{R}_+$ to be

$$\mathcal{B}_\lambda = \{\psi : \psi_x = \lambda \cdot \textstyle\prod_{k=1}^{p} \theta_{x_k},\ \theta_{x_k} \in [0,1],\ \text{for } x \in \mathcal{R}\} \tag{2.2}$$

and $\mathcal{B}_\infty = \lim_{\lambda \to \infty} B_\lambda$. A nonnegative CP decomposition is given by the summation $\psi = \sum_{k=1}^{\operatorname{rank}_+(\psi)} \psi^k$, where $\psi^k \in \mathcal{B}_\infty$ for $k \in [\operatorname{rank}_+(\psi)]$.

Convex Hull of Nonnegative Rank-1 Tensors

Now for $\lambda \in \mathbb{R}_+$ consider the finite set of points:

$$\mathcal{S}_\lambda = \{\psi : \psi_x = \lambda \cdot \textstyle\prod_{k=1}^{p} \theta_{x_k},\ \theta_{x_k} \in \{0,1\} \text{ for } x \in \mathcal{R}\}. \tag{2.3}$$

Using standard techniques from integer optimization [40, 77], the above set of points can be rewritten as a set of linear constraints on binary variables:

$$\begin{aligned}\mathcal{S}_\lambda = \{\psi : \lambda \cdot (1-p) + \lambda \cdot \textstyle\sum_{k=1}^{p} \theta_{x_k} \leq \psi_x,& \\ 0 \leq \psi_x \leq \lambda \cdot \theta_{x_k}, \theta_{x_k} \in \{0,1\},\ \text{for } k \in [p], x \in \mathcal{R}\}.&\end{aligned} \tag{2.4}$$

Our first result is that the convex hulls of the set of points $\mathcal{S}_\lambda$ and of the ball $\mathcal{B}_\lambda$ are equivalent.

Proposition 1. *We have the relation that* $\mathcal{C}_\lambda := \text{conv}(\mathcal{B}_\lambda) = \text{conv}(\mathcal{S}_\lambda)$.

Proof. We prove this by showing the two set inclusions $\text{conv}(\mathcal{S}_\lambda) \subseteq \text{conv}(\mathcal{B}_\lambda)$ and $\text{conv}(\mathcal{B}_\lambda) \subseteq \text{conv}(\mathcal{S}_\lambda)$. The first inclusion is immediate since by definition we have $\mathcal{S}_\lambda \subset \mathcal{B}_\lambda$, and so we focus on proving the second inclusion. We prove this by contradiction:

Suppose $\text{conv}(B_\lambda) \not\subset \text{conv}(\mathcal{S}_\lambda)$. Then there exists a tensor $\psi' \in B_\lambda$ with $\psi' \notin \text{conv}(\mathcal{S}_\lambda)$. By the hyperplane separation theorem, there exists $\varphi \in \mathbb{R}^{r_1 \times \cdots \times r_p}$ and $\delta > 0$ such that $\langle \varphi, \psi' \rangle \geq \langle \varphi, \psi \rangle + \delta$ for all $\psi \in \text{conv}(\mathcal{S}_\lambda)$, where $\langle \cdot, \cdot \rangle$ is the usual inner product that is defined as the summation of elementwise multiplication. Now consider the multilinear optimization problem

$$\begin{aligned} \max\ & \langle \varphi, \psi \rangle & & \\ \text{s.t. } & \psi_x = \lambda \cdot \textstyle\prod_{k=1}^{p} \theta_{x_k}, & & \text{for } x \in \mathcal{R} \\ & \theta_{x_k} \in [0,1], & & \text{for } x \in \mathcal{R} \end{aligned} \tag{2.5}$$

Proposition 2.1 of [28] shows that there exists a global optimum ψ'' of (2.5) such that $\psi'' \in \mathcal{S}_\lambda$. By construction of this global optimum, we must have $\langle \varphi, \psi'' \rangle \geq \langle \varphi, \psi' \rangle$, which implies $\langle \varphi, \psi'' \rangle \geq \langle \varphi, \psi \rangle + \delta$ for all $\psi \in \text{conv}(\mathcal{S}_\lambda)$. But this last statement is a contradiction since $\psi'' \in \mathcal{S}_\lambda \subseteq \text{conv}(\mathcal{S}_\lambda)$. ■

Remark 1. This result has two implications that will be important in later sections. The first is that $\mathcal{C}_\lambda$ is a polytope since it is the convex hull of a finite number of bounded points. The second is that the elements of $\mathcal{S}_\lambda$ are the vertices of $\mathcal{C}_\lambda$, since any individual element cannot be written as a convex combination of the other elements.

We will call the set $\mathcal{C}_\lambda$ the nonnegative tensor polytope. A useful observation is that the following relationships hold: $\mathcal{B}_\lambda = \lambda \mathcal{B}_1$, $\mathcal{S}_\lambda = \lambda \mathcal{S}_1$, and $\mathcal{C}_\lambda = \lambda \mathcal{C}_1$.

Constructing a Norm for Nonnegative Tensors

Since the set of nonnegative tensors forms a cone [82], we need to use a modified definition of a norm. A norm on a cone $\mathcal{K}$ is a function $p : \mathcal{K} \to \mathbb{R}_+$ such that for all $x, y \in \mathcal{K}$ the function has the following three properties: $p(x) = 0$ if and only if $x = 0$; $p(\gamma \cdot x) = \gamma \cdot p(x)$ for $\gamma \in \mathbb{R}_+$; and $p(x + y) \leq p(x) + p(y)$. The difference with the usual norm definition is

subtle: The second property here is required to hold for $\gamma \in \mathbb{R}_+$, whereas in the usual norm definition we require $p(\gamma \cdot x) = |\gamma| \cdot p(x)\ \forall \gamma \in \mathbb{R}$.

We next use $\mathcal{C}_\lambda$ to construct a new norm for nonnegative tensors using a gauge (or Minkowski functional) construction. Though constructing norms using a gauge is common in machine learning [23], the convex sets used in these constructions are symmetric about the origin. Symmetry guarantees that scaling the ball eventually includes the entire space [86]. However, in our case $\mathcal{C}_\lambda$ is not symmetric about the origin, and so without proof we do not *a priori* know whether scaling $\mathcal{C}_1$ eventually includes the entire space of nonnegative tensors. Thus we have to explicitly prove the gauge is a norm.

Proposition 2. *The function defined as*

$$\|\psi\|_+ := \inf\{\lambda \geq 0 \mid \psi \in \lambda \mathcal{C}_1\} \tag{2.6}$$

is a norm for nonnegative tensors $\psi \in \mathbb{R}_+^{r_1 \times \cdots \times r_p}$.

Proof. We first prove the above function is finite. Consider any nonnegative tensor $\psi \in \mathbb{R}_+^{r_1 \times \cdots r_p}$, and note there exists a decomposition $\psi = \sum_{i=1}^{\pi} \psi^k$ with $\psi^k \in \|\psi\|_{\max} \cdot \mathcal{B}_1$ [83]. (This holds because we can choose each ψ^k to have a single non-zero value that corresponds to a different entry of ψ.) Hence by Proposition 1 it follows that $\psi^k \in \|\psi\|_{\max} \cdot \mathcal{C}_1$. Recalling the decomposition of ψ, this means that $\psi \in \pi \|\psi\|_{\max} \cdot \mathcal{C}_1$ which by definition means $\|\psi\|_+ \leq \pi \|\psi\|_{\max}$. Thus $\|\psi\|_+$ must be finite.

Next we verify that the three properties of a norm on a cone are satisfied. To do so, we first observe that that by definition we have: $\mathcal{C}_1$ is convex; $0 \in \mathcal{C}_1$; and $\mathcal{C}_1$ is closed and bounded. Thus by Example 3.50 of [86] we have $\{\psi : \|\psi\|_+ = 0\} = \{0\}$, which means the first norm property holds. Also by Example 3.50 of [86] we have $\lambda \mathcal{C}_1 \subseteq \lambda' \mathcal{C}_1$ for all $0 \leq \lambda \leq \lambda'$, which means the second norm property holds. Last, we note Example 3.50 of [86] establishes sublinearity of the gauge, and so the third norm property holds for our case. ■

Convex Surrogate for Nonnegative Tensor Rank

An important property of our norm $\|\psi\|_+$ is that it can be used as a convex surrogate for tensor rank, whereas the max and Frobenius norms cannot.

Proposition 3. *If* ψ *is a nonnegative tensor, then we have*

$$\|\psi\|_{\max} \leq \|\psi\|_+ \leq \operatorname{rank}_+(\psi) \cdot \|\psi\|_{\max}. \tag{2.7}$$

If $\operatorname{rank}_+(\psi) = 1$*, then* $\|\psi\|_+ = \|\psi\|_{\max}$

Proof. Consider any $\lambda \geq 0$. If $\psi \in \mathcal{S}_\lambda$, then by definition $\|\psi\|_{\max} = \lambda$. By the convexity of norms we have that: if $\psi \in \mathcal{C}_\lambda$, then $\|\psi\|_{\max} \leq \lambda$. This means that for all $\lambda \geq 0$ we have $\mathcal{C}_\lambda \subseteq \mathcal{U}_\lambda := \{\psi : \|\psi\|_{\max} \leq \lambda\}$. Thus we have the relation

$$\inf\{\lambda \geq 0 \mid \psi \in \lambda \mathcal{U}_1\} \leq \inf\{\lambda \geq 0 \mid \psi \in \lambda \mathcal{C}_1\}. \tag{2.8}$$

But the left side is $\|\psi\|_{\max}$ and the right side is $\|\psi\|_+$. This proves the left side of the inequality (2.7).

Next consider the case where $\text{rank}_+(\psi) = 1$. By definition we have $\psi \in \|\psi\|_{\max} \cdot \mathcal{B}_1$. Thus $\psi \in \|\psi\|_{\max} \cdot \mathcal{C}_1$, and so $\|\psi\|_+ \leq \|\psi\|_{\max}$. This proves the rank-1 case.

Last consider any nonnegative tensor ψ. Recall its nonnegative CP decomposition $\psi = \sum_{k=1}^{\text{rank}_+(\psi)} \psi^k$ with $\psi^k \in \mathcal{B}_\infty$. Applying the triangle inequality gives

$$\|\psi\|_+ \leq \sum_{k=1}^{\text{rank}_+(\psi)} \|\psi^k\|_+ = \sum_{k=1}^{\text{rank}_+(\psi)} \|\psi^k\|_{\max}, \tag{2.9}$$

where the equality follows from the above-proved rank-1 case since $\text{rank}_+(\psi^k) = 1$ by definition of the CP decomposition. But $\|\psi^k\|_{\max} \leq \|\psi\|_{\max}$ since the tensors are nonnegative. This proves the right side of (2.7). ■

Remark 2. These bounds are tight. The lower and upper bounds are achieved by all nonnegative rank-1 tensors. As another example, the identity matrix with k columns achieves the upper bound with $\text{rank}_+(\psi) = k$.

We can prove a similar result for the Frobenius norm.

Proposition 4. *If ψ is a nonnegative tensor, then we have*

$$\tfrac{1}{\sqrt{\pi}} \cdot \|\psi\|_F \leq \|\psi\|_+ \leq \sqrt{\text{rank}_+(\psi)} \cdot \|\psi\|_F. \tag{2.10}$$

Proof. First note $\|\psi\|_{\max} \leq \|\psi\|_F \leq \sqrt{\pi}\|\psi\|_{\max}$, because these norms are elementwise and are hence subject to the bounds applicable to vector norms.

The left side of the inequality (2.10) now directly follows from the above inequalities and Proposition 3.

Next recall (2.9) from the proof of Proposition 3. Observe that $\|v\|_1 \leq \sqrt{d}\|v\|_2$ for a vector $v \in \mathbb{R}^d$. Thus we have

$$\begin{aligned} \|\psi\|_+ &\leq \sum_{k=1}^{\text{rank}_+(\psi)} \|\psi^k\|_{\max} \\ &\leq \sqrt{\text{rank}_+(\psi)} \cdot \Big(\sum_{i=k}^{\text{rank}_+(\psi)} (\|\psi^k\|_{\max})^2 \Big)^{1/2}. \end{aligned} \tag{2.11}$$

Let $x^k = \arg\max_{x \in \mathcal{R}} \psi^k{}_x$, and define $\{u^k\}$ to be the set of unique values from the set $\{x^k\}$. Since the tensors are nonnegative, we have $\psi_{u^k} \geq \sum_{k:x^k=u^k} \|\psi^k\|_{\max}$. Thus, we have that $\big(\sum_{k=1}^{\text{rank}_+(\psi)} (\|\psi^i\|_{\max})^2\big)^{1/2} \leq \big(\sum_{u^k} (\psi_{u^k})^2\big)^{1/2} \leq \|\psi\|_F$. The result follows by combining with (2.11). ■

Remark 3. These bounds are tight. The lower bound is achieved for a tensor with all entries equal to 1. The upper bound with $\text{rank}_+(\psi) = k$ is achieved for the identity matrix with k columns.

2.4 Measures of Norm Complexity

Computational Complexity

We next characterize the computational complexity of our new norm on nonnegative tensors.

Proposition 5. *It is* NP*-hard to approximate the nonnegative tensor norm $\|\cdot\|_+$ to arbitrary accuracy.*

Proof. The dual norm is $\|\varphi\|_\circ = \sup\{\langle\varphi,\psi\rangle \mid \|\psi\|_+ \leq 1\}$ for all tensors $\varphi \in \mathbb{R}^{r_1\times\cdots\times r_p}$. Theorems 3 and 10 of [34] show that approximation of the dual norm $\|\cdot\|_\circ$ is polynomial-time reducible to approximation of the norm $\|\cdot\|_+$. We proceed by showing a polynomial-time reduction to approximation of $\|\cdot\|_\circ$. Without loss of generality assume $p = 2$ and $d := r_1 = r_2$. The appendix of [109] proves that MAX CUT is polynomial-time reducible to $\sup\{\langle\varphi,\psi\rangle \mid \psi \in \mathcal{S}_1\}$. However, since the objective function is linear and the set $\mathcal{S}_1$ consists of the vertices of $\mathcal{C}_1$, this means that optimization problem is equivalent to $\sup\{\langle\varphi,\psi\rangle \mid \psi \in \mathcal{C}_1\}$. This is the desired polynomial-time reduction of MAX CUT to approximation of the dual norm $\|\cdot\|_\circ$ because $\mathcal{C}_1 = \{\psi : \|\psi\|_+ \leq 1\}$. The result now follows by recalling that approximately solving MAX CUT to arbitrary accuracy is NP-hard [78, 4]. ■

Corollary 4. *Given some $K \in \mathbb{R}_+$ and $\psi \in \mathbb{R}_+^{r_1\times\cdots\times r_p}$, it is* NP*-complete to determine whether $\|\psi\|_+ \leq K$.*

Proof. The problem is NP-hard by reduction from the problem of Proposition 5. In particular, a binary search can approximate the norm using this decision problem, $\|\psi\|_+ \overset{?}{\leq} K$, as an oracle. Applying Proposition 3, the search can be initialized over the interval $[0, \pi\|\psi\|_{\max}]$. Furthermore, the decision problem is in NP because we can use the (polynomial-sized) θ from (2.2) as a certificate. ■

Stochastic Complexity of Norm

We next show that our norm $\|\psi\|_+$ has low stochastic complexity. Let $X = \{x\langle 1\rangle, \ldots, x\langle n\rangle\}$, and suppose σ_i are independent and identically distributed (i.i.d.) Rademacher random variables (i.e., $\sigma_i = \pm 1$ with probability $\frac{1}{2}$) [10, 99]. The Rademacher complexity for a set of functions $\mathcal{H}$ is

$$\mathsf{R}(\mathcal{H}) = \mathbb{E}\Big(\sup_{h\in\mathcal{H}} \tfrac{1}{n}\big|\textstyle\sum_{i=1}^n \sigma_i \cdot h(x\langle i\rangle)\big|\Big), \tag{2.12}$$

and the *worst case* Rademacher complexity of $\mathcal{H}$ is

$$\mathsf{W}(\mathcal{H}) = \sup_X \mathbb{E}_\sigma\Big(\sup_{h\in\mathcal{H}} \tfrac{1}{n}\big|\textstyle\sum_{i=1}^n \sigma_i \cdot h(x\langle i\rangle)\big|\Big). \tag{2.13}$$

These notions can be used to measure the complexity of sets of matrices [98] or tensors [6]. The idea is to interpret each nonnegative tensor as a function $\psi : \mathcal{R} \to \mathbb{R}_+$ from a set of indices $x \in \mathcal{R}$ to the corresponding entry of the tensor ψ_x. This complexity notion is useful for the completion problem because it can be directly translated into generalization bounds.

Proposition 6. *We have $\mathsf{R}(\mathcal{C}_\lambda) \leq \mathsf{W}(\mathcal{C}_\lambda) \leq 2\lambda\sqrt{\rho/n}$.*

Proof. First note that from their definitions we get $\mathsf{R}(\mathcal{C}_\lambda) \leq \mathsf{W}(\mathcal{C}_\lambda)$. Next define the set $\mathcal{P}_\lambda = \{\pm\psi : \psi \in \mathcal{S}_\lambda\}$, and recall that $\mathcal{C}_\lambda = \operatorname{conv}(\mathcal{S}_\lambda)$ by Proposition 1. This means

$\mathsf{W}(\mathcal{C}_\lambda) = \mathsf{W}(\mathcal{S}_\lambda)$ [58, 10]. Next observe that

$$\begin{aligned}\mathsf{W}(\mathcal{C}_\lambda) = \mathsf{W}(\mathcal{S}_\lambda) &= \sup_X \mathbb{E}_\sigma \Big(\sup_{\psi \in \mathcal{S}_\lambda} \tfrac{1}{n} \big| \textstyle\sum_{i=1}^n \sigma_i \cdot \psi_{x\langle i \rangle} \big| \Big) \\ &= \sup_X \mathbb{E}_\sigma \Big(\max_{\psi \in \mathcal{P}_\lambda} \tfrac{1}{n} \cdot \textstyle\sum_{i=1}^n \sigma_i \cdot \psi_{x\langle i \rangle} \Big) \\ &\leq \sup_X r \sqrt{2 \log \#\mathcal{P}_\lambda}/n \end{aligned} \tag{2.14}$$

where in the second line we have replaced the supremum with a maximum since the set $\mathcal{P}_\lambda$ is finite and in the last line we have used the Finite Class Lemma [67] with

$$r = \max_{\psi \in \mathcal{P}_\lambda} \sqrt{\textstyle\sum_{i=1}^n (\psi_{x\langle i \rangle})^2} \leq \lambda \sqrt{n} \tag{2.15}$$

This inequality on r is due to the fact that $\mathcal{P}_\lambda$ consists of tensors whose entries are from $\{-\lambda, 0, \lambda\}$. Thus $\mathsf{W}(\mathcal{C}_\lambda) \leq \lambda \sqrt{(2 \log 2) \cdot (\rho + 1)/n}$. The result follows by weakening the bound using $\log 2 \cdot (\rho + 1) \leq 2\rho$. ■

Remark 5. Because ψ has $\pi = O(r^p)$ entries, the Rademacher complexity in a typical tensor norm (e.g., max and Frobenius norms) will be $O(\sqrt{\pi/n}) = O(\sqrt{r^p/n})$. However, Rademacher complexity in our norm $\|\psi\|_+$ is $O(\sqrt{\rho/n}) = O(\sqrt{rp/n})$, which is exponentially smaller.

2.5 Tensor Completion

Suppose we have data $(x\langle i \rangle, y\langle i \rangle) \in \mathcal{R} \times \mathbb{R}$ for $i = 1, \ldots, n$. Let $I = \{i_1, \ldots, i_u\} \subseteq [n]$ be any set of points that specify all the unique $x\langle i \rangle$ for $i = 1, \ldots, n$, meaning the set $U = \{x\langle i_1 \rangle, \ldots, x\langle i_u \rangle\}$ does not have any repeated points and $x\langle i \rangle \in U$ for all $i = 1, \ldots, n$. The nonnegative tensor completion problem using our norm $\|\psi\|_+$ is given by

$$\begin{aligned} \widehat{\psi} \in \arg\min_{\psi} \; & \tfrac{1}{n} \textstyle\sum_{i=1}^n \big(y\langle i \rangle - \psi_{x\langle i \rangle} \big)^2 \\ \text{s.t. } & \|\psi\|_+ \leq \lambda \end{aligned} \tag{2.16}$$

We will study the statistical properties of the above estimate and describe the elements of algorithm to solve the above optimization problem. The purpose of defining the set U is that it is used in constructing the algorithm used to solve the above problem.

Statistical Guarantees

Though in the previous section we calculated a Rademacher complexity for nonnegative tensors in $\mathcal{C}_\lambda$ viewed as functions, here we use an alternative approach to derive generalization bounds. The reason is that generalization bounds using the Rademacher complexity are not tight here.

Our approach is based on the observation that the nonnegative tensor completion problem (2.16) using our norm $\|\psi\|_+$ is equivalent to a *convex aggregation* problem [74, 102, 57] for a finite set of functions. In particular, by Proposition 1 we have $\{\psi : \|\psi\|_+ \leq \lambda\} = \mathcal{C}_\lambda = \operatorname{conv}(\mathcal{S}_\lambda)$. The implication is we can directly apply existing results for convex aggregation to provide a tight generalization bound for the solution of (2.16).

Proposition 7 ([57]). *Suppose $|y| \leq b$ almost surely. Given any $\delta > 0$, with probability at least $1 - 4\delta$ we have that*

$$\mathbb{E}\big((y - \psi_x)^2\big) \leq \min_{\varphi \in \mathcal{C}_\lambda} \mathbb{E}\big((y - \varphi_x)^2\big) + c_0 \cdot \max\big[b^2, \lambda^2\big] \cdot \max\Big[\zeta_n, \tfrac{\log(1/\delta)}{n}\Big], \tag{2.17}$$

where

$$\zeta_n = \begin{cases} \frac{2^\rho}{n}, & \text{if } 2^\rho \leq \sqrt{n} \\ \sqrt{\frac{1}{n} \log\left(\frac{e 2^\rho}{\sqrt{n}}\right)}, & \text{if } 2^\rho > \sqrt{n} \end{cases} \tag{2.18}$$

and c_0 is an absolute constant.

Remark 6. We make two comments. First, note that $\zeta_n = o(\sqrt{\rho/n})$. Second, in some regimes ζ_n can be considerably faster than the $\sqrt{\rho/n}$ rate.

Generalization bounds under specific noise models, such as an additive noise model, follow as a corollary to the above proposition combined with Proposition 3.

Corollary 7. *Suppose φ is a nonnegative tensor with $\operatorname{rank}_+(\varphi) = k$ and $\|\varphi\|_{\max} \leq \mu$. If $(x\langle i\rangle, y\langle i\rangle)$ are independent and identically distributed with $|y\langle i\rangle - \varphi_{x\langle i\rangle}| \leq e$ almost surely and $\mathbb{E}y\langle i\rangle = \varphi_{x\langle i\rangle}$. Then given any $\delta > 0$, with probability at least $1 - 4\delta$ we have*

$$\mathbb{E}\big((y - \widehat{\psi}_x)^2\big) \leq e^2 + c_0 \cdot \big(\mu k + e\big)^2 \cdot \max\Big[\zeta_n, \tfrac{\log(1/\delta)}{n}\Big], \tag{2.19}$$

where ζ_n is as in (2.18) and c_0 is an absolute constant.

Computational Complexity

Though (2.16) is a convex optimization problem, our next result shows that solving it is NP-hard.

Proposition 8. *It is NP-hard to solve the nonnegative tensor completion problem (2.16) to an arbitrary accuracy.*

Proof. Define the ball of radius $\delta > 0$ centered at a nonnegative tensor ψ to be

$$B(\psi, \delta) = \big\{\varphi : \|\varphi - \psi\|_F \leq \delta\big\}. \tag{2.20}$$

Next define $W(\mathcal{C}_1, \delta) = \bigcup_{\psi \in \mathcal{C}_1} B(\psi, \delta)$ and $W(\mathcal{C}_1, -\delta) = \{\psi \in \mathcal{C}_1 : B(\psi, \delta) \subseteq \mathcal{C}_1\}$. The weak membership problem for $\mathcal{C}_1$ is that given a nonnegative tensor ψ and a $\delta > 0$ decide whether $\psi \in W(\mathcal{C}_1, \delta)$ or $\psi \notin W(\mathcal{C}_1, -\delta)$. Theorem 10 of [34] shows that approximation of the norm $\|\cdot\|_+$ is polynomial-time reducible to the weak membership problem for $\mathcal{C}_1$. Since Proposition 5 shows that approximation of the norm $\|\cdot\|_+$ is NP-hard, the result follows if we can reduce the weak membership problem to (2.16).

Suppose we are given inputs ψ and δ for the weak membership problem. Choose $x\langle i\rangle$ for $i = 1, \ldots, \pi$ such that each element in $\mathcal{R}$ is enumerated exactly once. Next choose $y\langle i\rangle = \psi_{x\langle i\rangle}$. Finally, note if we solve (2.16) and the minimum objective value is less than or equal to δ, then we can conclude $\psi \in W(\mathcal{C}_1, \delta)$; otherwise we have $\psi \notin W(\mathcal{C}_1, -\delta)$. The result now follows since this was the desired reduction. ∎

We note that the decision version of (2.16), that is ascertaining whether a given tensor ψ attains an objective less than a given value while also satisfying $\|\psi\|_+ \leq \lambda$, is NP-complete. This follows directly from Corollary 4 and Proposition 8.

Numerical Computation

Although it is NP-hard, there is substantial structure that enables efficient numerical computation of global minima of (2.16). The key observation is that $\mathcal{C}_1$ is a 0-1 polytope, which implies the linear separation problem on this polytope can be solved using integer linear optimization. Integer optimization has well-established global algorithms that are guaranteed to solve the separation problem for this polytope. This is a critical feature that enables the use of the Frank-Wolfe algorithm or one of its variants to solve (2.16) to a desired numerical tolerance. In fact, the Frank-Wolfe variant known as *Blended Conditional Gradients* (BCG) [17] is particularly well-suited for calculating a solution to (2.16) for the following reasons:

The first reason is that our problem has structure such that the BCG algorithm will terminate (within numerical tolerance) in a linear number of oracle steps. A sufficient condition for this linear convergence is if the feasible set is a polytope and the objective function is strictly convex over the feasible set. For our problem (2.16), the objective function can be made strictly convex over the feasible set by an equivalent reformulation. Specifically, we use the equivalent reformulation in which we change the feasible set from $\{\psi : \|\psi\|_+ \leq \lambda\} = C_\lambda$ to $\text{Proj}_U(\mathcal{C}_\lambda)$ where the projection is done over the unique indices specified by the set U. In fact, this projection is trivial to implement because it simply consists of discarding the entries of ψ that are not observed by the $x\langle i\rangle$ indices.

The second is that the BCG algorithm uses weak linear separation, which accommodates early-termination of the associated integer linear optimization. Integer optimization software tends to find optimal or near-optimal solutions considerably faster than certifying the optimality of such solutions. Furthermore, a variety of tuning parameters (see e.g. [13]) can be used to accelerate the search for good primal solutions when exact solutions are not needed. We also deploy a fast alternating maximization heuristic in order to avoid integer optimization oracle calls where possible. Thus, early-termination allows us to deploy a globally convergent algorithm with practical solution times.

Hence we consider the BCG algorithm to compute global minima of (2.16) to arbitrary numerical tolerance. For brevity we omit a full description of the algorithm, and instead focus on the separation oracle, which is the main component specifically adapted to our application. The oracle is described in Algorithm 1, where we use $\langle \cdot, \cdot \rangle$ to denote the dot product for tensors viewed as vectors; note that for notational simplicity we state the oracle in terms of the original space, ignoring projection onto U. Output condition (1) provides separation with some vertex φ, while (2) requires certification that no such vertex exists.

In implementing the weak separation oracle, we first attempt separation with our alternating maximization procedure. Described as Algorithm 2, it involves the following objective function:

$$z_M(\theta) := \sum_{x \in \mathcal{R}} \langle c_x, \psi_x - \lambda \cdot \prod_{k=1}^{p} \theta_{x_k} \rangle \tag{2.21}$$

The separation problem is thus treated as a multilinear binary optimization problem, and the algorithm successively minimizes in each dimension. We apply this heuristic a fixed number

of times, randomly complementing entries in the incumbent each time. The procedure runs in polynomial-time, and is not guaranteed to separate nor can it certify that such separation is impossible. However, in our experiments it succeeds often, offering substantial practical speedups.

If alternating maximization fails to separate, then we solve the following integer programming problem:

$$\begin{aligned} &\max_{\varphi,\theta} \langle c, \psi - \varphi \rangle \\ &\text{s.t. } \lambda \cdot (1-p) + \lambda \cdot \textstyle\sum_{k=1}^{p} \theta_{x_k} \leq \varphi_x && x \in \mathcal{R} \\ &\qquad 0 \leq \varphi_x \leq \lambda \cdot \theta_{x_k} && k \in [p], x \in \mathcal{R} \\ &\qquad \theta_{x_k} \in \{0,1\} && k \in [p], x \in \mathcal{R} \end{aligned} \tag{2.22}$$

Note that early termination is deployed: we stop whenever an incumbent solution is found such that the objective is greater than Φ/K. If no such solution exists, then the integer programming solver is guaranteed to (eventually) establish a dual bound z^U such that $\langle c, \psi - \varphi \rangle \leq z^U \leq \Phi$.

Algorithm 1 Weak Separation Oracle for $\mathcal{C}_\lambda$

Input: linear objective $c \in \mathbb{R}^{r_1 \times \cdots \times r_p}$, point $\psi \in \mathcal{C}_\lambda$, accuracy $K \geq 1$, gap estimate $\Phi > 0$, norm bound λ

Output: Either (1) vertex $\varphi \in \mathcal{S}_\lambda$ with $\langle c, \psi - \varphi \rangle \geq \Phi/K$, or (2) **false:** $\langle c, \psi - \varphi \rangle \leq \Phi$ for all $\varphi \in \mathcal{C}_\lambda$

Algorithm 2 Alternating Maximization

Input: linear objective $c \in \mathbb{R}^{r_1 \times \cdots \times r_p}$, point $\psi \in \mathcal{C}_\lambda$, norm bound λ, incumbent (binary) solution $\hat{\theta} \in \mathcal{S}_\lambda$

Output: Best known solution θ

$\theta \leftarrow \hat{\theta}$
$z \leftarrow z_M(\theta)$
for $i = 1$ **to** p **do**
  for $k = 1$ **to** r_i **do**
    $\theta_k^{(i)} \leftarrow 1 - \theta_k^{(i)}$
    if $z_M(\theta) > z$ **then**
      $z \leftarrow z_M(\theta)$
    else
      $\theta_k^{(i)} \leftarrow 1 - \theta_k^{(i)}$
    end if
  end for
end for

Table 2.1: Numerical Results with Order-3 Nonnegative Tensors (n=500)

Tensor Size	**BCG with Weak Oracle**		**Alternating Least Squares**	
	NMSE	Time (s)	NMSE	Time (s)
$10 \times 10 \times 10$	0.004 ± 0.001	4.29 ± 0.194	0.058 ± 0.002	0.122 ± 0.003
$20 \times 20 \times 20$	0.070 ± 0.003	28.6 ± 0.753	0.174 ± 0.005	0.351 ± 0.012
$30 \times 30 \times 30$	0.220 ± 0.004	11.0 ± 0.671	0.287 ± 0.005	0.796 ± 0.030
$40 \times 40 \times 40$	0.300 ± 0.004	5.83 ± 0.202	0.351 ± 0.005	0.947 ± 0.043
$50 \times 50 \times 50$	0.371 ± 0.004	4.86 ± 0.090	0.406 ± 0.006	1.58 ± 0.093
$60 \times 60 \times 60$	0.434 ± 0.004	4.78 ± 0.063	0.454 ± 0.006	2.08 ± 0.105
$70 \times 70 \times 70$	0.500 ± 0.004	4.94 ± 0.063	0.503 ± 0.008	3.30 ± 0.140
$80 \times 80 \times 80$	0.554 ± 0.004	5.73 ± 0.112	0.539 ± 0.010	5.96 ± 0.226
$90 \times 90 \times 90$	0.594 ± 0.004	5.34 ± 0.077	0.576 ± 0.011	9.34 ± 0.302
$100 \times 100 \times 100$	0.641 ± 0.004	5.56 ± 0.072	0.666 ± 0.013	16.8 ± 0.568

Table 2.2: Numerical Results with Dimension-10 Nonnegative Tensors (n=10,000)

Tensor Size	**BCG with Weak Oracle**		**Alternating Least Squares**	
	NMSE	Time (s)	NMSE	Time (s)
$10^{\times 4}$	0.001 ± 0.000	25.2 ± 1.11	0.007 ± 0.001	0.925 ± 0.025
$10^{\times 5}$	0.002 ± 0.000	40.7 ± 0.82	0.074 ± 0.006	1.13 ± 0.038
$10^{\times 6}$	0.008 ± 0.001	59.5 ± 1.20	0.412 ± 0.025	1.79 ± 0.097
$10^{\times 7}$	0.034 ± 0.004	530 ± 368	0.930 ± 0.019	1.28 ± 0.050
$10^{\times 8}$	0.156 ± 0.014	872 ± 66	0.986 ± 0.010	15.1 ± 0.199

Table 2.3: Numerical Results with Tensor Size $10^{\times 6}$ and Increasing Sample Size

Sample %	**BCG with Weak Oracle**		**Alternating Least Squares**	
	NMSE	Time (s)	NMSE	Time (s)
0.01	0.989 ± 0.002	0.348 ± 0.016	1.00 ± 0.00	0.067 ± 0.001
0.1	0.457 ± 0.023	62.4 ± 4.79	1.00 ± 0.00	$0.105 \pm .001$
1	0.007 ± 0.000	60.4 ± 1.32	0.389 ± 0.023	1.82 ± 0.098
10 ($n = 100,000$)	0.005 ± 0.000	413.7 ± 11.1	0.046 ± 0.005	15.03 ± 0.401

2.6 Numerical Experiments

Here we present numerical results that show the efficacy and scalability of our algorithm for nonnegative tensor completion. Our experiments were conducted on a personal computer with 8GB of RAM and an Intel Core i5 2.3Ghz processor with 2-cores/4-threads. The algorithms were coded in Python 3. We used Gurobi v9.1 [39] to solve the integer programs (2.22).

As a benchmark, we use alternating least squares (ALS), which is often called a "workhorse" for numerical tensor problems [52]. We believe ALS alone is the correct benchmark because it is substantially more scalable than other algorithms, such as those using semidefinite programming (SDP) [71, 9, 85]. In fact, many of the problems we consider in our experiments (the largest tensors have 10^8 entries) are too large for nearly all tensor completion algorithms that have been previously developed.

To minimize the impact of hyperparameter selection in our numerical results, we provided the ground truth values when possible. For instance, in our nonnegative tensor completion formulation (2.16) we chose λ to be the smallest value for which we could certify that $\|\psi\|_+ \leq \lambda$ for the true tensor ψ. This was accomplished by construction of the true tensor ψ. For ALS, we used a nonnegative rank k that was the smallest value for which we could certify that $\text{rank}_+(\psi) \leq k$. This was also accomplished by construction of the true tensor ψ. A last note is that ALS often works better when used with L2 regularization [73]. The hyperparameter for the L2 regularization for ALS was chosen in a way favorable to ALS in order to maximize its accuracy.

Experiments with Third-Order Tensors

Our first set of results concerns tensors of order $p = 3$ with increasing dimension values. In each experiment, the true tensor ψ was constructed by randomly choosing 10 points from the set of points $\mathcal{S}_1$ and then taking a random convex combination of these 10 points. This method of construction of ψ ensures that $\|\psi\|_+ \leq 1$ and $\text{rank}_+(\psi) \leq 10$. We used $n = 500$ samples (with indices sampled with replacement). Each experiment was repeated 100 times, and the results are shown in Table 2.1 in terms of "normalized mean squared error $\pm$ standard error". The measure of accuracy we used is normalized mean squared error (NMSE) $\|\widehat{\psi} - \psi\|_F^2 / \|\psi\|_F^2$. This is a more stringent measure of accuracy than used in Corollary 7 because the statistical theoretical result does not include normalization. The numerical results in the table show that, on average, our approach provides modestly higher accuracy but requires more computation time. However, the computation time remains on the order of seconds for the various tensor sizes. For reference, by construction of the 0-1 polytope, the worst-case normalized mean squared error is 1, implying no meaningful result was found, while the best case NMSE is 0, implying an exact construction of the underlying true tensor. We also note that NMSE is not exactly zero because indices are sampled with replacement and because of the finite precision of numerical algorithms.

Experiments with Increasing Tensor Order

Our second set of results concerns tensors with increasing order p, where each dimension takes the value $r_i = 10$ for $i = 1, \ldots, p$. In each experiment, the true tensor was constructed by the

method given in Section 5.1, again ensuring that $\|\psi\|_+ \leq 1$ and $\text{rank}_+(\psi) \leq 10$. We used $n = 10,000$ samples (with indices sampled with replacement). Each experiment was repeated 100 times, and the results are shown in Table 2.2 in terms of "NMSE $\pm$ standard error". The numerical results in the table show that our approach provides substantially higher accuracy but requires more computation time. However, the computation time remains on the order of minutes even for the largest tensor with 10^8 entries.

Experiments with Increasing Sample Size

Our third and fourth set of results concerns tensors of size $10^{\times 6}$ and $10^{\times 7}$, respectively, where experiments are conducted with increasing sample size, given in the table as percentage of total entries. For example, for $n = 100$ on a 10^6 tensor, the sample percentage is 100/1,000,000 $= 0.01\%$. The true tensor ψ was constructed as in Section 5.1, with the same assurances. We begin with sample percentage 0.01% (with indices sampled with replacement), and extend each set of experiments' sample size by one order of magnitude. Each experiment was repeated 100 times, and the results are shown in Table 2.3 and Table 2.4 in terms of "NMSE $\pm$ standard error", as above. Again, the numerical results in the table show that our approach provides substantially higher accuracy but requires more computation time. The computation time remains on the order of minutes for most of the sampling schemes, excepting the draw of 10^6 entries for a tensor with 10^7 entries (i.e., about 1.5 hours).

Table 2.4: Numerical Results with Tensor Size $10^{\times 7}$ and Increasing Sample Size

Sample %	**BCG with Weak Oracle**		**Alternating Least Squares**	
	NMSE	Time (s)	NMSE	Time (s)
0.01	0.873 ± 0.016	48.4 ± 3.42	1.00 ± 0.00	0.561 ± 0.003
0.1	0.029 ± 0.004	145 ± 36.6	0.877 ± 0.023	1.29 ± 0.043
1	0.011 ± 0.001	522 ± 11.7	0.220 ± 0.016	19.2 ± 0.603
10 ($n = 1,000,000$)	0.013 ± 0.001	5623 ± 381	0.052 ± 0.007	189 ± 4.89

Chapter 3

Sample Average Approximation

Sample Average Approximation (SAA) is used to approximately solve stochastic optimization problems. Consider a generic stochastic optimization problem of the form

$$\min_{x\in\mathcal{X}}\{F(x) := \mathbb{E}_\xi f(x,\xi)\}, \tag{3.1}$$

where $\xi \in \Xi$ is some random variable with known distribution $P(\cdot)$, and $\mathcal{X} \subset \mathbb{R}^p$ is the (deterministic) feasible set. That is, we consider problems where stochasticity enters into the objective function and not the constraints. Note $F(x^*) = \mathbb{E}_\xi f(x^*,\xi)$ is the corresponding optimal value, where

$$x^* \in \arg\min_{x\in\mathcal{X}} F(x)$$

is any optimal point. Such problems are often difficult to solve because the expectation cannot be analytically computed except in cases when the distribution $P(\cdot)$ or the function $f(x,\xi)$ have very specific mathematical forms.

Sample Average Approximation (SAA) is a commonly-used procedure for solving (3.1), and it works by approximating the stochastic optimization problem using a deterministic optimization problem that is easier to solve [29, 51, 103, 94, 50]. The idea of SAA is to first generate an i.i.d. sample $\xi_1,\ldots,\xi_n$ of the random variable ξ, and then approximate the expectation $\mathbb{E}_\xi f(x,\xi)$ using its sample average

$$\min_{x\in\mathcal{X}}\{F_n(x) := \frac{1}{n}\sum_{i=1}^{n} f(x,\xi_i)\}. \tag{3.2}$$

Note the objective function value of the original stochastic optimization problem (3.1) with the optimal solution of the SAA problem (3.2) is given by $F(\hat{x}_n) = \mathbb{E}_\xi f(\hat{x}_n,\xi)$, where we have that

$$\hat{x}_n \in \arg\min_{x\in\mathcal{X}} F_n(x)$$

is any optimal point of the SAA problem (3.2). By definition of x^* and $\hat{x}_n$, we have that $F(x^*) \le F(\hat{x}_n)$, and clearly we expect to see that $F(\hat{x}_n) \to F(x^*)$ almost surely as $n \to \infty$ by an argument using the uniform law of large numbers.

3.1 Classical Sample Bounds

Two practical considerations necessitate that the number of samples n in the SAA problem (3.2) be as small as possible. First, for many applications it is computationally costly to generate any single sample ξ_i of the random variable ξ. Second, for many functional forms of $f(\cdot, \cdot)$ it is the case that larger values of n require greater computation (e.g., more function evaluations, more gradient evaluations, etc.) in order to numerically solve the SAA problem (3.2).

Towards this goal, a now classical analysis [51, 93, 94] showed that in order to ensure

$$\mathbb{P}\big(F(\hat{x}_n) - F(x^*) \leq \delta\big) \geq 1 - \alpha \tag{3.3}$$

for any $\delta \in (0, 1]$ and $\alpha \in (0, 1]$, the number of samples n should satisfy

$$n \gtrsim \frac{p}{\delta^2} \log \frac{1}{\delta} + \frac{1}{\delta^2} \log \frac{1}{\alpha}. \tag{3.4}$$

Here, we have used the notation $x \gtrsim y$ of [63] which means $x \geq cy$ for some constant $c > 0$ that is independent of p, δ, α and which may depend polynomially upon other parameters of the optimization problem (3.1). This bound says the required number of samples depends polynomially on the dimension p of the decision variable x, and this bound is prohibitively restrictive when p is high-dimensional. In fact, applications of SAA for large p are common in many domains.

However, the experience of many practitioners has been that a much smaller number of samples n (as compared to the above bound) is needed in order to ensure the SAA approximation is close to the true optimal value [90, 65, 50] – that is, the sample bound (3.4) is often overly conservative. Motivated by this empirical observation, there has been work on algorithmic approaches that iteratively or adaptively choose sample sizes in order to get good solutions with SAA with a small number of samples [89, 88].

3.2 Logarithmic Bounds with Special Structure

We briefly outline our chapter and highlight our main contributions. We first provide an overview of the stochastic process theory of Rademacher complexity [59, 11]. Our first contribution uses stochastic process theory to derive a new bound for n to ensure (3.3). This bound depends in a non-trivial way upon the complex, stochastic interplay between $\mathcal{X}$ and $f(\cdot, \cdot)$. In the next section, we describe an algorithmic procedure that can be used to numerically upper-bound this stochastic quantity for some stochastic optimization problems (3.1). Then in the following section, we give examples of stochastic optimization problems where our approach yields explicit symbolic bounds on the number of samples n needed. Notably, we show that single-index models with an ℓ_1 constraint yields logarithmic bounds on the number of samples needed. We conclude by conducting numerical experiments with the Markowitz portfolio selection problem to demonstrate the significant improvement of our bound relative to the classical bound (3.4).

Comparison to Other Sample Bounds

One set of previous results [51, 93, 94] are based on an intermediate bound: When $\mathcal{X}$ is a discrete set of points, then (3.3) whenever $n \gtrsim \frac{1}{\delta^2} \log \frac{\#\mathcal{X}}{\alpha}$, where $\#\mathcal{X}$ is the cardinality of the finite set $\mathcal{X}$. Such a counting approach is effective for some problems: For example, these ideas have been used to prove that nonconvex (both integer and continuous) optimization problems with an ℓ_1 constraint have a significantly reduced theoretical computational complexity [69].

However, the above counting bound obscures the complex interplay between the feasible set $\mathcal{X}$ and the mathematical structure of the function $f(\cdot, \cdot)$, which is what actually governs the behavior of the SAA solutions. More recent work [75, 76] uses empirical process theory to derive sample bounds, which is better able to capture the interplay between the objective and the feasible set. This work uses a chaining argument to characterize stochastic complexity, and this approach is in fact closely related our use of Rademacher theory that also characterizes stochastic complexity.

All sample bounds make assumptions about the continuity of $f(x, \xi)$ and about the problem stochasticity. It is common to assume $f(x, \xi)$ satisfies stochastic Lipschitz [51, 93, 94, 37] or stochastic Hölder continuity [75, 76] conditions. It is also common to assume stochastic regularity, such as assuming it follows sub-Gaussian distributions [51, 93, 94, 37] or assuming that sample averages are well-behaved. These results find sample bounds that are linear in the dimension p [51, 93, 94, 37], or find sample bounds for specific problems like lasso that have effective dimensions that are smaller than p (and even logarithmic in p in the case of lasso) [75, 76].

In this chapter, we assume deterministic Lipschitz continuity on $f(x, \xi)$ and ensure stochastic regularity by requiring boundedness. We show sample bounds that are logarithmic in dimension p when the underlying stochastic optimization problem has ℓ_1 or nuclear norm constraints. The sample bounds of [51, 93, 94, 37] are linear in p because they consider generic feasible sets, whereas the faster-than-lienar bounds of [75, 76] are calculated only for specific problems. However, we stress that our logarithmic sample bounds do not arise because we have made stronger assumptions, but they are instead due to the geometry of the ℓ_1 or nuclear norm constraints. We demonstrate this by providing in this paper an alternative proof of logarithmic bounds under the more general assumptions of [51, 93, 94]. Our stronger assumptions are due to the proof technique we use. For instance, our boundedness assumption can be relaxed using the results of [53] though we do not consider this generalization here because it requires more notation that hides the main ideas.

Another related line of work has explored sample bounds under assumptions of sparsity. The work in [63, 62] used sparsity-based techniques to study the relationship between sample size and SAA solution quality for the special case where the optimal solution x^* is sparse. (Here we define sparsity to include both low cardinality vectors and low rank matrices.) The approach of [63, 62] adds a regularization term to the SAA (3.2), which induces sparsity in the optimal solution $\hat{x}$ to the SAA. The authors prove that this leads the sample size n to have poly-logarithmic dependence on the dimension p. However, there is an alternative explanation for these derived bounds. Since the optimal solution is known to be sparse, the effective feasible set $\mathcal{X}'$ (which incorporates the fact that x^* is sparse) can be taken to be much smaller than the stated feasible set $\mathcal{X}$. We use this idea in Sec. 3.5 to derive similar logarithmic bounds for a similar (to the ones studied by [63, 62]) class of problems; these

original bounds were achieved using regularization and using a much more technically difficult argument than the one we provide here.

3.3 Rademacher Complexities

Before deriving our results, we need to provide a brief introduction to the stochastic process theory of Rademacher complexity [59, 11], which is key to understanding our results. We provide this introduction because these results and underlying methodology are not generally known within the operations and control communities, though they are generally well-known within statistics and probability theory.

Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher random variables, where ϵ is a Rademacher random variable if its distribution is $\mathbb{P}(\epsilon = \pm 1) = \frac{1}{2}$; and let $f(x, \xi)$ be the function from the objective of (3.1). We define the *Rademacher complexity* of the function set $\mathcal{F} := \{f(x, \xi) : x \in \mathcal{X}\}$ to be

$$\mathcal{R}_n[f] = \mathbb{E}_\xi\Big(\sup_{x \in \mathcal{X}} \Big| \tfrac{1}{n} \textstyle\sum_{i=1}^n \epsilon_i f(x, \xi_i) \Big| \Big). \tag{3.5}$$

Note the Rademacher complexity is often defined without an absolute value when the set $\mathcal{F}$ is symmetric, and we can use an equivalent definition without an absolute value

$$\mathcal{R}_n[f] = \mathbb{E}_\xi\Big(\sup_{s \in \pm 1, x \in \mathcal{X}} \tfrac{1}{n} \textstyle\sum_{i=1}^n \epsilon_i s f(x, \xi_i) \Big)$$

by defining the augmented function set $\mathcal{F}' := \{sf(x, \xi) : s \in \pm 1, x \in \mathcal{X}\}$. Because both definitions are equivalent, we use the representation (3.5) to maintain consistency. Without loss of generality, we will make the following assumption for the remainder of the paper:

Assumption. *We have that* $-\Delta/2 \leq f(x, \xi) \leq \Delta/2$ *for all* $(x, \xi) \in \mathcal{X} \times \Xi$*, for some finite constant* $\Delta \in \mathbb{R}_+$*.*

This assumption is without loss of generality because if the function $f(x, \xi)$ is bounded on its domain $(x, \xi) \in \mathcal{X} \times \Xi$ then we can always define an equivalent stochastic optimization problem by defining $f'(x, \xi) = f(x, \xi) - m - \Delta/2$ for some finite constant $m \in \mathbb{R}$ such that the above is satisfied.

We would like to emphasize that the above boundedness assumption can be relaxed by using recent generalizations of McDiarmid's inequality for unbounded random variables [53], but we do not consider this generalization here because it adds considerable notational complexity while obscuring the underlying ideas. In this paper, we assume boundedness so as to most clearly focus on the key underlying ideas.

The Rademacher complexity can be used to construct inequalities that bound the probability of large deviations of certain stochastic processes from their expectations. We begin with such a result on deviation between the sample average (3.2) from the objective of SAA with its expectation (3.1) in the objective of the stochastic optimization problem. The result below is similar to existing ones on concentration of measure [59, 11, 16], though our result differs somewhat from the standard forms. Its proof uses what is known as a *symmetrization* argument, and we repeat this argument here to introduce it to readers who are unfamiliar with it.

Proposition 9. *If the assumption holds, then we have*

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}\Big|F_n(x)-F(x)\Big|>t\Big)\leq\exp\Big(-2n\Big(\frac{t-2\mathcal{R}_n[f]}{\Delta}\Big)^2\Big). \tag{3.6}$$

Proof. For notational convenience, we define

$$\mathcal{E}=\sup_{x\in\mathcal{X}}\Big|\tfrac{1}{n}\textstyle\sum_{i=1}^n f(x,\xi_i)-\mathbb{E}_\xi f(x,\xi)\Big|$$

Observe that Jensen's inequality gives

$$\mathbb{E}\Big(\mathcal{E}\Big)\leq\mathbb{E}\Big(\sup_{x\in\mathcal{X}}\Big|\tfrac{1}{n}\textstyle\sum_{i=1}^n f(x,\xi_i)-f(x,\xi_i')\Big|\Big),$$

where ξ_i,ξ_i' are i.i.d. But $f(x,\xi_i)-f(x,\xi_i')$ has a symmetric distribution, and so its distribution is equivalent to the distribution of $\epsilon_i\cdot(f(x,\xi_i)-f(x,\xi_i'))$ since the ϵ_i have a Rademacher distribution. And so we get

$$\mathbb{E}\Big(\mathcal{E}\Big)\leq\mathbb{E}\Big(\sup_{x\in\mathcal{X}}\Big|\tfrac{1}{n}\textstyle\sum_{i=1}^n \epsilon_i(f(x,\xi_i)-f(x,\xi_i'))\Big|\Big).$$

Applying the triangle inequality yields

$$\mathbb{E}\Big(\mathcal{E}\Big)\leq 2\mathbb{E}\Big(\sup_{x\in\mathcal{X}}\Big|\tfrac{1}{n}\textstyle\sum_{i=1}^n \epsilon_i f(x,\xi_i)\Big|\Big). \tag{3.7}$$

But observe that the right-hand side is $2\mathcal{R}_n[f]$. Since the function $f(\cdot,\cdot)$ is bounded by assumption, we can use the standard McDiarmid's inequality [16, 68] to get

$$\mathbb{P}\Big(\mathcal{E}-\mathbb{E}\Big(\mathcal{E}\Big)>u\Big)\leq\exp\Big(-2n\Big(\frac{u}{\Delta}\Big)^2\Big).$$

Combining this with (3.7) implies that

$$\mathbb{P}\Big(\mathcal{E}-2\mathcal{R}_n[f]>u\Big)\leq\exp\Big(-2n\Big(\frac{u}{\Delta}\Big)^2\Big),$$

or with the substitution $u=t-2\mathcal{R}_n[f]$ that (3.6) holds. ∎

We also prove a nonstandard result (i.e., to the best of our knowledge this result is not found in the literature on Rademacher complexity) involving functions of (3.1).

Corollary 1. *Let $h:\mathbb{R}\to\mathbb{R}$ be Lipschitz with constant L, and suppose the assumption holds. Then we have*

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}\Big|h(F_n(x))-h(F(x))\Big|>t\Big)\leq\exp\Big(-2n\Big(\frac{t-2L\mathcal{R}_n[f]}{L\Delta}\Big)^2\Big). \tag{3.8}$$

Proof. We first note that Lipschitz continuity of $h(\cdot)$ implies $|h(F_n(x))-h(F(x))|\leq L|F_n(x)-F(x)|$. This means by Proposition 9 we have

$$\begin{aligned}\mathbb{P}\Big(\sup_{x\in\mathcal{X}}\Big|h(F_n(x))-h(F(x))\Big|>t\Big)&\leq\mathbb{P}\Big(\sup_{x\in\mathcal{X}}L\Big|F_n(x)-F(x)\Big|>t\Big)\\&\leq\exp\Big(-2n\Big(\frac{t/L-2\mathcal{R}_n[f]}{\Delta}\Big)^2\Big).\end{aligned} \tag{3.9}$$

The bound (3.8) now follows. ∎

It is pragmatically useful to interpret this corollary. In essence, the result says that applying a Lipschitz function $h(\cdot)$ to (3.1) is the same in terms of the concentration of measure as scaling the Rademacher complexity by L to $L\mathcal{R}_n[f]$ and scaling the assumption bound by L to $L\Delta$.

3.4 New Sample Bounds

The above proposition can be used to construct new sample bounds that ensure (3.3) holds for a generic stochastic optimization problem (3.1), and our next result gives an implicit formula for such a bound when $\mathcal{R}_n[f]$ is strictly decreasing.

Proposition 10. *Suppose $R_n[f]$ is strictly decreasing in n. If the assumption holds and $n \geq N$ with*

$$\begin{aligned} N &= \min_{\gamma_1,\gamma_2} \max\{N_1, N_2\} \\ N_1 &= (\Delta/\gamma_1\delta)^2 \log(2/\alpha)/2 \\ N_2 &= \min\{n : R_n[f] \leq \gamma_2\delta\} \end{aligned} \tag{3.10}$$

and $\gamma_1, \gamma_2 \in (0,1)$ such that $2\gamma_1 + \gamma_2 = 1$, then (3.3) holds.

Proof. First observe that

$$F(\hat{x}_n) - F(x^*) = F(\hat{x}_n) - F_n(\hat{x}_n) + F_n(\hat{x}_n) - F_n(x^*) + F_n(x^*) - F(x^*). \tag{3.11}$$

Since $\hat{x}_n$ minimizes the SAA, we have $F_n(\hat{x}_n) \leq F_n(x^*)$. Let $\gamma_1, \gamma_2 \in (0,1)$ be such that $2\gamma_1 + \gamma_2 = 1$, and note the union bound gives

$$\begin{aligned} \mathbb{P}\Big(F(\hat{x}_n) - F(x^*) \leq \delta\Big) \geq 1 - \mathbb{P}\Big(F(\hat{x}_n) - F_n(\hat{x}_n) > (\gamma_1+\gamma_2)\delta\Big) + \\ - \mathbb{P}\Big(F_n(x^*) - F(x^*) > \gamma_1\delta\Big). \end{aligned} \tag{3.12}$$

Next, observe that Hoeffding's inequality [42, 16] implies

$$\mathbb{P}\Big(F_n(x^*) - F(x^*) > t\Big) \leq \exp\Big(-2N\Big(\frac{t}{\Delta}\Big)^2\Big)$$

since $n \geq N$. Note we use Hoeffding's inequality (i.e., not a uniform convergence result) since x^* is fixed. So if $-2N \cdot (\gamma_1\delta/\Delta)^2 = \log(\alpha/2)$ then $\mathbb{P}(F_n(x^*) - F(x^*) > \gamma_1\delta) \leq \alpha/2$. Now if $\gamma_2\delta \geq 2\mathcal{R}_n[f]$, then by Proposition 9 we have

$$\mathbb{P}\Big(F(\hat{x}_n) - F_n(\hat{x}_n) > (\gamma_1+\gamma_2)\delta\Big) \leq \exp\Big(-2n\Big(\frac{\gamma_1\delta}{\Delta}\Big)^2\Big).$$

Since $n \geq N$, then if $-2N \cdot (\gamma_1\delta/\Delta)^2 = \log(\alpha/2)$ then $\mathbb{P}(F(\hat{x}_n) - F_n(\hat{x}_n) > (\gamma_1+\gamma_2)\delta) \leq \alpha/2$. Combing the above with (3.12) gives the desired result.

■

The above result can be difficult to interpret because of the implicit equation that specifies the minimum sample size N for (3.3) to hold. So we next present a simplified result for the case where the Rademacher complexity can be bounded in the form $\mathcal{R}_n[f] \leq \frac{c(p)}{\sqrt{n}}$ for some function $c(p)$. Such a bound can be constructed for many cases [59, 46, 92, 48, 47].

Corollary 2. *If the assumption holds,* $\mathcal{R}_n[f] = \frac{c(p)}{\sqrt{n}}$, *and*

$$n \geq \frac{1}{2}\left(\frac{3\Delta}{\delta}\right)^2 \log\left(\frac{2}{\alpha}\right) + \left(\frac{3c(p)}{\delta}\right)^2$$

then we have that (3.3) holds.

Proof. We first compute $N_2 = \min\{n : R_n[f] \leq \gamma_2\delta\}$ under the assumptions of this corrollary. Specifically, we have $c(p)/\sqrt{N_2} = \gamma_2\delta$, or rewritten that $N_2 = (c(p)/\gamma_2\delta)^2$. Next consider an $N' := N_1 + N_2 \geq N = \min_{\gamma_1,\gamma_2} \max\{N_1, N_2\}$. Noting that $\gamma_2 = 1 - 2\gamma_1$, we have that

$$N' = (\Delta/\gamma_1\delta)^2 \log(2/\alpha)/2 + (c(p)/(1-2\gamma_1)\delta)^2.$$

Our result follows by choosing $\gamma_1 = 1/3$.

■

3.5 Bounds for Problems

The prior two results bound the sample size n required to achieve (3.3), but their use requires knowing the Rademacher complexity $\mathcal{R}_n[f]$ for a particular stochastic optimization problem (3.1). Here, we discuss how the Rademacher complexity can be bounded for various classes of stochastic optimization problems. We describe a Monte Carlo approach, and we also give explicit bounds for specific problems.

Monte Carlo Bounds

The first class of problems we consider are those where there exists a surrogate optimization problem that provides an upper bound of the form

$$\sup_{x\in\mathcal{X}} \left| \tfrac{1}{n} \textstyle\sum_{i=1}^n \epsilon_i f(x, \xi_i) \right| \leq \max_{y\in\mathcal{Y}} G(y, \{\epsilon_i, \xi_i\}_{i=1}^n). \tag{3.13}$$

Our scenario is where the optimization problem on the right-hand side is easily-solvable. For example, when the problem on the left-hand side can be represented as a mixed-integer linear program (MILP) then the problem on the right-hand side could be its continuous relaxation. Various approximation techniques exist, and how different classes of problems for the left-hand side can be upper-bounded by a surrogate problem are beyond the scope of this present paper.

For this scenario, we define a Monte Carlo estimate. Let ξ_{ij} and ϵ_{ij} be i.i.d. samples of ξ and of the Rademacher random variable, respectively. Then a Monte Carlo estimate $\widehat{\mathcal{R}}[f]$ of the Rademacher complexity is given by

$$\widehat{\mathcal{R}}_n[f] = \frac{1}{m} \sum_{j=1}^m \left(\max_{y_j\in\mathcal{Y}} G(y_j, \{\epsilon_{ij}, \xi_{ij}\}_{i=1}^n) \right)$$

where m is the number of repetitions. Our next result concerns the correctness of this Monte Carlo estimate.

Proposition 11. *Suppose* $0 \leq G(y, \{\epsilon_i, \xi_i\}^n) \leq \frac{\sigma}{2}$ *for all* $(y, \{\epsilon_i, \xi_i\}^n) \in \mathcal{Y} \times \{\pm 1, \Xi\}^n$*, for some finite constant* $\sigma \in \mathbb{R}_+$*. Then we have* $\mathcal{R}_n[f] \leq \widehat{\mathcal{R}}_n[f] + d$ *with probability at least* $1 - \exp(-2m(\frac{d}{\sigma})^2)$.

Proof. First note that Hoeffding's inequality implies

$$\mathbb{P}\Big(\widehat{\mathcal{R}}_n[f] < \mathbb{E}\Big(\widehat{\mathcal{R}}_n[f]\Big) - d\Big) \leq \exp\Big(-2m\Big(\frac{d}{\sigma}\Big)^2\Big).$$

We next define the quantity

$$\mathcal{F} = \tfrac{1}{m}\textstyle\sum_{j=1}^m \Big(\sup_{x_j \in \mathcal{X}} \Big|\tfrac{1}{n}\sum_{i=1}^n \epsilon_{i,j} f(x_j, \xi_{i,j})\Big|\Big).$$

But by definition $\mathcal{F} \leq \widehat{\mathcal{R}}_n[f]$ and $\mathbb{E}\mathcal{F} = \mathcal{R}_n[f] \leq \mathbb{E}\widehat{\mathcal{R}}_n[f]$, and so this means that we have

$$\mathbb{P}\Big(\widehat{\mathcal{R}}_n[f] < \mathcal{R}_n[f] - d\Big) \leq \exp\Big(-2m\Big(\frac{d}{\sigma}\Big)^2\Big).$$

The result follows by taking the complement of this. ■

The above result says the Monte Carlo estimate $\widehat{\mathcal{R}}_n[f]$ upper bounds the Rademacher complexity $\mathcal{R}_n[f]$ with high probability. The quality of the estimate depends upon two factors. The first is how weak the upper bound (3.13) of the surrogate optimization problem is. The second is how large m is, with larger values corresponding to more accurate estimates. We reiterate that this approach is only feasible when the surrogate problem is easily-solvable, which enables the use of large values of m in computing the estimate. We will provide a specific example in the next section.

Explicit Bounds

Next, we provide explicit bounds for specific classes of stochastic optimization problems. Continuity is needed for logarithmic bounds even with favorable feasible sets. For instance, Proposition 2 of [37] gives an example $f(x, \xi)$ with $\mathcal{X} = \{x : \|x\|_2 \leq \lambda\}$, where $F(x)$ is Lipschitz and linear bounds are necessary for a small optimality gap. By defining $g(x)$ such that $g(0) = 0$ and $g(x) = x \cdot \|x\|_1/\|x\|_2$ otherwise, we thus construct an example $f(g(x), \xi)$ with $\mathcal{X} = \{x : \|x\|_1 \leq \lambda\}$ where $F(g(x))$ is not Lipschitz and a linear number of samples is necessary for a small optimality gap.

Proposition 12. *Let* $g : \mathbb{R} \to \mathbb{R}$ *be Lipschitz with constant* L*, and consider the stochastic optimization problem*

$$\min_{x \in \mathcal{S}} \Big\{\mathbb{E}_\xi\big(g(\xi^\mathsf{T} x)\big) \;\Big|\; \|x\|_1 \leq \lambda\Big\} \tag{3.14}$$

where $\mathcal{S} \subseteq \mathbb{R}^p$ *and* $\max_{\xi \in \Xi} \|\xi\|_\infty \leq C < +\infty$*. Then the Rademacher complexity of the above problem is bounded by* $\mathcal{R}_n[f] \leq \lambda LC\sqrt{2\log 2p/n}$*, and we need*

$$n \geq \Big(\frac{3\lambda LC}{\delta}\Big)^2 \cdot \Big(2\log\Big(\frac{2}{\alpha}\Big) + 2\log 2p\Big) \tag{3.15}$$

samples to ensure that (3.3) holds.

Proof. Let $\Lambda = \{x : \|x\|_1 \leq \lambda\}$, and note that we have

$$\max_{x,y\in\Lambda} \left|g(\xi^\mathsf{T} x) - g(\xi^\mathsf{T} y)\right| \leq L \max_{x,y\in\Lambda} \left|\xi^\mathsf{T}(x-y)\right| \leq 2\lambda LC$$

where the first inequality follows by Lipschitz continuity of $g(\cdot)$, and the second inequality follows by Hölder's inequality. This means the assumption holds for $\Delta = 2\lambda LC$. Next we bound the Rademacher complexity of the problem (3.14), which is bounded by L times the Rademacher complexity for the stochastic optimization problem where $g(\cdot)$ is the identity function (see Lemma 26.9 in [92], which was originally Lemma 1.1 from the lecture notes [46]; a slightly less general version of this result was first shown by [59]). This second Rademacher complexity for when $g(\cdot)$ is the identity was bounded in [48], and the final result is as above. The sample bound (3.15) now follows from Corollary 2. ■

The above single-index model with an ℓ_1 constraint is a situation where we need logarithmic in p samples for SAA, which is a substantial improvement over the standard bound (3.4) showed by [51, 93, 94] that is polynomial in p.

Next we consider a class of problems similar to [63], and we show a similar logarithmic in p bound but without using regularization and with a much simpler technical argument.

Corollary 3. *Let $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz with constant L, and consider the stochastic optimization problem*

$$\min_{x\in\mathcal{X}} \mathbb{E}_\xi\big(g(\xi^\mathsf{T} x)\big) \tag{3.16}$$

where $\max_{\xi\in\Xi} \|\xi\|_\infty \leq C < +\infty$. Suppose there is an optimal solution x^* to (3.16) that is sparse, meaning that $s := \sum_{i=1}^p \mathbf{1}(x_i^* \neq 0)$ is small with $\|x^*\|_\infty \leq \mu < +\infty$. Then

$$n \geq \left(\frac{3\mu s LC}{\delta}\right)^2 \cdot \left(2\log\left(\frac{2}{\alpha}\right) + 2\log 2p\right)$$

samples ensures that (3.3) holds when $\hat{x}$ is the SAA solution to the stochastic optimization problem

$$\min_{x\in\mathcal{X}} \left\{\mathbb{E}_\xi\big(g(\xi^\mathsf{T} x)\big) \;\middle|\; \|x\|_1 \leq \mu s\right\}. \tag{3.17}$$

Proof. Note that $\|x^*\|_1 \leq \mu s$ by assumption. Thus x^* is an optimal solution for both (3.17) and (3.16), and both problems have the same minimum value $F(x^*)$. The result now follows by applying Proposition 12 to (3.17). ■

Last we note that our sample bound can be improved when the problem has nonegativity constraints.

Proposition 13. *Let $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz with constant L, and consider the stochastic optimization problem*

$$\min_{x\in\mathcal{S}} \left\{\mathbb{E}_\xi\big(g(\xi^\mathsf{T} x)\big) \;\middle|\; x \geq 0, \|x\|_1 \leq \lambda\right\}$$

where $\mathcal{S} \subseteq \mathbb{R}^p$ and $\max_{\xi\in\Xi} \|\xi\|_\infty \leq C < +\infty$. Then the Rademacher complexity of the above problem is bounded by $\mathcal{R}_n[f] \leq \lambda LC\sqrt{\log p/n}$, and we need

$$n \geq \left(\frac{3\lambda LC}{\delta}\right)^2 \cdot \left(\frac{1}{2}\log\left(\frac{2}{\alpha}\right) + \log p\right)$$

samples to ensure that (3.3) holds.

The proof is omitted because it is essentially identical to that of Proposition 12, the main difference being a different bound from [48] is used for the Rademacher complexity.

Explicit Bounds for Matrix Optimization

Next, we provide bounds for specific classes of stochastic matrix optimization problems. We use $\|X\|_*$ to denote the nuclear norm of $X \in \mathbb{R}^{p\times q}$, and $\|\xi\|_2$ in this section denotes the spectral norm of the random matrix $\xi \in \Xi \subset \mathbb{R}^{p\times q}$.

Proposition 14. *Let $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz with constant L, and consider the stochastic optimization problem*

$$\min_{X\in\mathcal{S}} \left\{ \mathbb{E}_\xi\big(g(\mathsf{tr}(\xi^\mathsf{T} X))\big) \;\middle|\; \|X\|_* \le \lambda \right\}$$

where $\mathcal{S} \subseteq \mathbb{R}^{p\times q}$ and $\max_{\xi\in\Xi} \|\xi\|_2 \le C < +\infty$. Then the Rademacher complexity of the above stochastic optimization problem is bounded by $\mathcal{R}_n[f] \le \lambda LC\sqrt{3\log(\min\{p,q\})/n}$, and we need

$$n \ge \left(\frac{3\lambda LC}{\delta}\right)^2 \cdot \left(2\log\left(\frac{2}{\alpha}\right) + 3\log\big(\min\{p,q\}\big)\right)$$

samples to ensure that (3.3) holds.

Proof. Let $\Lambda = \{X : \|X\|_* \le \lambda\}$, and note that we have

$$\begin{aligned}
\max_{X,Y\in\Lambda} \Big|g(\mathsf{tr}(\xi^\mathsf{T} X)) - g(\mathsf{tr}(\xi^\mathsf{T} Y))\Big| & \\
&\le L \max_{X,Y\in\Lambda} \Big|\mathsf{tr}(\xi^\mathsf{T}(X-Y))\Big| \\
&\le L \max_{X,Y\in\Lambda} \|\xi\|_2 \|X-Y\|_* \\
&\le 2\lambda LC
\end{aligned}$$

where the first inequality follows by Lipschitz continuity of $g(\cdot)$, and the second inequality follows by Hölder's inequality for unitarily invariant norms [15]. This means the assumption holds for $\Delta = 2\lambda LC$. Next we bound the Rademacher complexity of (3.14): Observe that this is bounded by L times the Rademacher complexity for when the function $g(\cdot)$ is the identity function (see Lemma 26.9 in [92] and Lemma 1.1 in [46]). The Rademacher complexity for when $g(\cdot)$ is the identity was bounded in [47], and the final result is as above. The sample bound (3.15) now follows from Corollary 2. ∎

The above single-index model with a nuclear norm constraint needs logarithmic in $\min\{p,q\}$ samples for SAA. This is a substantial improvement over the standard bound (3.4) showed by [51, 93, 94] that in this case is

$$n \gtrsim \frac{pq}{\delta^2}\log\frac{1}{\delta} + \frac{1}{\delta^2}\log\frac{1}{\alpha}.$$

which is polynomial in p and q.

Next we consider a class of problems similar to [62], and we show a logarithmic in $\min\{p,q\}$ bound but without using regularization and with a much simpler technical argument.

Corollary 4. *Let $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz with constant L, and consider the stochastic optimization problem*

$$\min_{X \in \mathcal{X}} \mathbb{E}_\xi\big(g(\mathsf{tr}(\xi^\mathsf{T} X))\big) \tag{3.18}$$

where $\mathcal{X} \subseteq \mathbb{R}^{p\times q}$ and $\max_{\xi\in\Xi} \|\xi\|_2 \le C < +\infty$. Suppose there is an optimal solution X^* to (3.18) that is low rank, meaning $r := \mathsf{rank}(X^*)$ is small with $\|X^*\|_2 \le \mu < +\infty$. Then we need

$$n \ge \Big(\frac{3\mu r LC}{\delta}\Big)^2 \cdot \Big(2\log\Big(\frac{2}{\alpha}\Big) + 3\log\big(\min\{p,q\}\big)\Big)$$

samples to ensure that (3.3) holds when $\hat{X}$ is the SAA solution to the stochastic optimization problem

$$\min_{x\in\mathcal{X}} \Big\{\mathbb{E}_\xi\big(g(\mathsf{tr}(\xi^\mathsf{T} X)) \;\Big|\; \|X\|_* \le \mu r\Big\}. \tag{3.19}$$

Proof. Note that $\|X^*\|_* \le \mu r$ by assumption. Thus X^* is an optimal solution for both (3.19) and (3.18), and both problems have the same minimum value $F(x^*)$. The result now follows by applying Proposition 14 to (3.19). ■

Alternative Proof

We conclude this section by considering the more general setting of the past bound from [94]. We give an alternative proof of our result for problems with an ℓ_1 constraint, which modifies the proof of Theorem 5.18 in [94].

Proposition 15. *Consider the stochastic optimization*

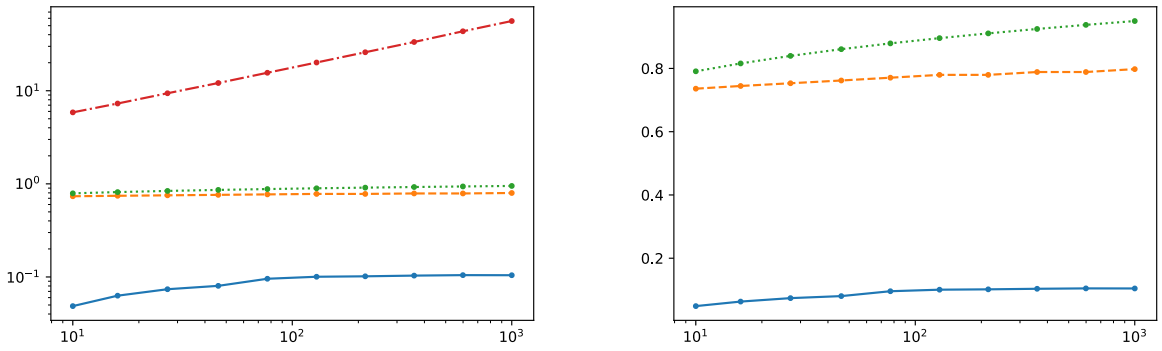
$$\min_{x\in\mathcal{S}} \Big\{\mathbb{E}_\xi f(x,\xi) \;\Big|\; \|x\|_1 \le \lambda\Big\} \tag{3.20}$$

where $\mathcal{S} \subseteq \mathbb{R}^p$. Let $\Lambda = \{x \in \mathcal{S} : \|x\|_1 \le \lambda\}$, and suppose two assumptions hold. First, for any $x', x \in \Lambda$ there exists constant $\sigma_{x',x} > 0$ such that the moment-generating function $M_{x',x}(t) = \mathbb{E}_\xi \exp(tY_{x',x})$ of random variable $Y_{x',x} = [f(x',\xi) - F(x')] - [f(x,\xi) - F(x)]$ satisfies $M_{x',x}(t) \le \exp(\sigma^2_{x',x}t^2/2)$ for all $t \in \mathbb{R}$. Second, there exists a (measurable) function $\kappa : \Xi \to \mathbb{R}_+$ such that its moment-generating function $M_\kappa(t)$ is finite valued for all t in a neighborhood of zero and $|f(x',\xi) - f(x,\xi)| \le \kappa(\xi)\|x' - x\|$ for almost everywhere $\xi \in \Xi$ and all $x', x \in \Lambda$. Then (3.3) holds whenever

$$n \ge \frac{8\sigma^2}{\delta^2} \cdot \log\frac{64\lambda^2 L^2 p}{\delta^2} + \Big(\frac{8\sigma^2}{\delta^2} + \frac{1}{\beta}\Big) \cdot \log\Big(\frac{2}{\alpha}\Big), \tag{3.21}$$

where $\sigma^2 = \sup_{x',x\in\Lambda}(\sigma_{x',x})^2$, $L = \mathbb{E}_\xi\kappa(\xi)$, and we have that $\beta = \sup_{t\in\mathbb{R}}\big(2Lt - \log M_\kappa(t)\big)$.

Proof. We first show there exists a set $\mathcal{V} = \{x_1, \ldots, x_k\}$ with $\log k \le 32(\lambda L/\delta)^2 \log p$ so $\max_{x\in\Lambda} \min_{x'\in\mathcal{X}} \|x - x'\| \le \delta/8L$. To show this, we use the Sudakov minoration [100] that says $\sqrt{\log k} \le \mathbb{E}(\sup_{x\in\Lambda} g^\mathsf{T} x)/2(\delta/8L)$, where $g \in \mathbb{R}^p$ is a vector whose entries are i.i.d. Gaussian random variables with zero mean and unit variance. Hölder's inequality and the symmetry of Λ imply that we have $\mathbb{E}(\sup_{x\in\Lambda} g^\mathsf{T} x) \le \mathbb{E}(\sup_{x\in\Lambda} \|x\|_1 \cdot \max_j |g_j|) \le \lambda\sqrt{2\log p}$ where we have used the basic bound $\mathbb{E}(\max_j |g_j|) \le \sqrt{2\log p}$ for g_j that are the j-th entry of



(a) log-log plot

(b) semi-log plot

Figure 3.1: Comparison of 95% upper confidence bound of SAA solution gap (solid blue) with bounds on 95% upper confidence bound gap predicted by [51, 93, 94] (dash-dotted red), our Proposition 10 (dashed orange), and our Corollary 2 (dotted green). The left shows results on a log-log scale, and the right shows results (excluding the [51, 93, 94] bound) on a semi-log scale. In both plots, the x-axis is the dimension p of the decision variable, and the y-axis is the 95% upper confidence bound gap.

the vector g. Thus Sudakov's minoration gives that $\sqrt{\log k} \leq \lambda\sqrt{2\log p}/2(\delta/8L)$. Rearranging this inequality gives the desired bound $\log k \leq 32(\lambda L/\delta)^2 \log p$.

Next choose any x^* that minimizes (3.20) and consider the modified stochastic optimization problem $\min_{x\in\mathcal{V}\cup\{x^*\}} F(x)$. Let $x^{*,v} \in \arg\min_{x\in\mathcal{V}\cup\{x^*\}} F(x)$, and define the solution $\hat{x}_n^v \in \arg\min_{x\in\mathcal{V}\cup\{x^*\}} F_n(x)$. Note $F_n(\hat{x}_n) \leq F_n(\hat{x}_n^v)$ since $\mathcal{V}\cup\{x^*\} \subseteq \Lambda$. The second assumption implies that with probability one we have $|F_n(x') - F_n(x)| \leq \hat{\kappa}_n\|x' - x\|$ for all $x', x \in \Lambda$, where $\hat{\kappa}_n = \frac{1}{n}\sum_{i=1}^n \kappa(\xi_i)$. By construction of $\mathcal{V}$, there exists x' with $\|x' - \hat{x}_n\| \leq \delta/8L$. Thus $F_n(x') \leq F_n(\hat{x}_n) + \hat{\kappa}_n \cdot (\delta/8L) \leq F_n(\hat{x}_n^v) + \hat{\kappa}_n \cdot (\delta/8L)$.

We continue our analysis under the event that $\hat{\kappa}_n \leq 2L$. Here, $F_n(x') \leq F_n(\hat{x}_n^v)+\delta/4$. Thus Theorem 5.17 of [94] says that $\mathbb{P}\big(F(x')-F(x^{*,v}) \leq 3\delta/4\big) \geq 1-\alpha/2$ for $n \geq (8\sigma^2/\delta^2)\times\log(2(k+1)/\alpha)$. But $F(x^{*,v}) = F(x^*)$ by construction of the modified stochastic optimization, and also $F(\hat{x}_n) \leq F(x')+L\cdot(\delta/8L)$ since the second assumption implies that $|F(x')-F(x)| \leq L\|x'-x\|$ for all $x', x \in \Lambda$. This means we have $F(\hat{x}_n) \leq F(x^*) + 3\delta/4 + \delta/8$ with probability at least $1-\alpha/2$. Note the Chernoff bound implies that $\hat{\kappa}_n \leq 2L$ with probability at least $1-\exp(-n\beta) = 1-\alpha/2$ when $n \geq \log(2/\alpha)/\beta$. Thus (3.3) holds when (3.21) holds. ■

The significance of this alternative proof is that it shows that the logarithmic rates arise because of the properties of ℓ_1 constraint. A similar alternative proof can also be constructed for nuclear norm constraints. The boundedness and Lipschitz continuity assumptions we make in previous subsections are artifacts of the proof technique that we use.

3.6 Numerical Experiments

Consider a scenario where we would like to choose a portfolio that allocates investments into some combination of p risky assets and 1 risk-free asset, while considering a tradeoff between maximizing the expected return of the portfolio and the risk tolerance of the investor. The Markowitz portfolio selection model [66, 19] is a simple framework to pose such a problem. Let $\xi \in \mathbb{R}^p$ is a random variable of the returns from the p risky assets, and define $\mu = \mathbb{E}_\xi \xi$ and $\Sigma = \mathbb{E}_\xi((\xi - \mu)(\xi - \mu)^\mathsf{T})$. Then one formulation of the problem involves solving a convex quadratic program

$$\min_{x \in \mathbb{R}^p} \left\{ x^\mathsf{T} \Sigma x - \gamma \cdot x^\mathsf{T}(\mu - r\mathbf{1}) \;\middle|\; x \geq 0, \|x\|_1 \leq 1 \right\} \tag{3.22}$$

where: r is the rate of return for the risk-free asset, $\gamma > 0$ trades-off betewen the returns and risk of the portfolio, and each entry of the vector x gives the fraction of the portfolio allocated to the p risky assets; hence $1 - \sum_{i=1}^p x_i$ is the fraction of the portfolio allocated to the risk-free asset.

Sample Bounds

To bound the Rademacher complexity of (3.22), we can use an existing calculus for Rademacher complexity [59, 11].

Proposition 16. *If $\|\xi\|_\infty \leq s$, then the Rademacher complexity for (3.22) is bounded by $\mathcal{R}_n[f] \leq (4s^2 + \gamma s)\sqrt{\log p / n}$. Also, the assumption is satisfied for $\Delta = 4s^2 + \gamma s$.*

Proof. Using the identity $\Sigma = \mathbb{E}_\xi(\xi\xi^\mathsf{T}) - \mu\mu^\mathsf{T}$, we have $x^\mathsf{T}\Sigma x - \gamma \cdot (\mu - r\mathbf{1})^\mathsf{T} x = \mathbb{E}_\xi\big((\xi^\mathsf{T}x)^2 - \gamma \cdot \xi^\mathsf{T}x\big) - \big(\mathbb{E}_\xi\big(\xi^\mathsf{T}x\big)\big)^2 + \gamma \cdot r\mathbf{1}^\mathsf{T}x$. Thus we can rewrite (3.22) as

$$\begin{aligned} &\min \mathbb{E}_\xi\big((\xi^\mathsf{T}x)^2 - \gamma \cdot \xi^\mathsf{T}x\big) - \big(\mathbb{E}_\xi\big(\xi^\mathsf{T}x\big)\big)^2 + \gamma \cdot r\mathbf{1}^\mathsf{T}x \\ &\text{s.t. } x \geq 0, \|x\|_1 \leq 1 \end{aligned}$$

The above is useful for bounding Rademacher complexity. Deterministic terms have a Rademacher complexity of zero, and the Rademacher complexity for the sum of problems is upper-bounded by the sum of the individual Rademacher complexities [59, 11]. So we conclude the proof by bounding the Rademacher complexity of three problems: First, consider the problem $\min\{\mathbb{E}_\xi((\xi^\mathsf{T}x)^2) \mid x \geq 0, \|x\|_1 \leq 1\}$. Since $\|\xi\|_\infty \leq s$, Proposition 13 says $\mathcal{R}_n[f_1] \leq 2s^2\sqrt{\log p/n}$ with $\Delta_1 = 2s^2$, since $g(u) = u^2$ is Lipschitz with $L = 2s$ when $u \in [-s, s]$. Second, consider the optimization problem $\min\{-(\mathbb{E}_\xi(\xi^\mathsf{T}x))^2 \mid x \geq 0, \|x\|_1 \leq 1\}$. Proposition 13 with Corollary 1 gives $R_n[f_2] \leq 2s^2\sqrt{\log p/n}$ with $\Delta_2 = 2s^2$, since $h(u) = u^2$ is Lipschitz with $L = 2s$ when $u \in [-s, s]$. Third, consider the problem $\min\{\mathbb{E}_\xi(-\gamma \cdot \xi^\mathsf{T}x) \mid x \geq 0, \|x\|_1 \leq 1\}$. Proposition 13 gives $R_n[f_3] \leq \gamma s\sqrt{\log p/n}$ with $\Delta_3 = \gamma s$. The result follows by noting that $\mathcal{R}_n[f] \leq \mathcal{R}_n[f_1] + \mathcal{R}_n[f_2] + \mathcal{R}_n[f_3]$ and that $\Delta \leq \Delta_1 + \Delta_2 + \Delta_3$. ■

We can compare various sample bounds on n to ensure (3.3) holds. We begin by calculating the specific previous bound from [94]: Hoeffding's lemma [42] bounds variance of (3.22) by

$\frac{\Delta^2}{4}$ for the Δ from Proposition 16. Moreover, the Lipschitz constant of the objective (without expectation) in (3.22) is $L \leq \sqrt{p}(4s^2 + \gamma s)$. Consequently, the previous bound from [94] is

$$n \geq 2 \cdot \left(\frac{4s^2 + \gamma s}{\delta}\right)^2 \cdot \left(p \log \frac{8\sqrt{p}(4s^2 + \gamma s)}{\delta} + \log \frac{2}{\alpha}\right). \tag{3.23}$$

In contrast, combining Proposition 16 with Corollary 2 gives our bound to be

$$n \geq \left(\frac{12s^2 + 3\gamma s}{\delta}\right)^2 \cdot \left(\frac{1}{2} \log \left(\frac{2}{\alpha}\right) + \log p\right). \tag{3.24}$$

The difference is our bound (3.24) is logarithmic in p whereas the past bound (3.23) is quasi-linear in p.

Results of Experiment

To experimentally compare the above bounds with the actual SAA performance, we consider this next scenario: We assume returns for the assets are distributed as

$$\xi \sim \mathcal{U}^1(-\tfrac{s}{2}, \tfrac{s}{2})\mathbf{1}_p + \mathcal{U}^p(-\tfrac{s}{2}, \tfrac{s}{2}),$$

where $\mathcal{U}^k(l, u)$ is a k-dimensional uniform distribution with the support of the distribution in each dimension of $[l, u]$, and $\mathbf{1}_p$ is a p-dimensional vector of ones. The interpretation is that the first term $\mathcal{U}^1(-\frac{s}{2}, \frac{s}{2})\mathbf{1}_p$ describes a strongly correlated component of the returns, while the second term $\mathcal{U}^p(-\frac{s}{2}, \frac{s}{2})$ describes an independent component of the returns. We also assume $\gamma = 1$, $s = 1$, and $r = 0$.

To compare the bounds with the actual SAA solution gaps, we used $n = 50$ and computed the 95% upper confidence bound of the SAA solution gap by solving SAA a total of 10,000 times each for different values of p. The 95% upper confidence bound is the smallest value of δ for $\alpha = 0.05$ in (3.3). We also compute the smallest value of δ for $\alpha = 0.05$ for the different bounds available to us. Specifically, we compare the actual upper confidence bound to

- bound (3.23), which is the past bound from [51, 93, 94]
- bound (3.10), which is the implicit sample bound from our Proposition 10
- bound (3.24), which is the simplified sample bound from our Corollary 2 for when $\lim_n \delta = 0$

The results are shown in Fig. 3.1. The past bound from [51, 93, 94] grows much faster than the actual SAA solution gap. In contrast, our bound (3.10) from Proposition 10 visually matches the growth rate of the actual SAA solution gap. Our bound (3.24), which is the simplified bound using Corollary 2, grows grows faster than the actual SAA solution gap; however, the bound (3.24) is a reasonably accurate approximation to (3.10) from Proposition 10. This suggests the simplified bound of Corollary 2 is useful for qualitative understanding of scaling, whereas the more accurate bound of Proposition 10 is more useful for determining necessary sample sizes for SAA.

Chapter 4

A Philosophical Framework for OIT

Following the presentation of stochastic models, theory, and analysis techniques in the last two chapters, we now move to propose a framework for OIT, based in stochastic optimization and modelling. This chapter provides the philosophical framework for optimal intervention theory, and is based in economic theory, with special attention paid to behavioral psychology and other fields that contribute to the study of optimal incentive and contract design. OIT is the study of population *adoption* of *socially beneficial goods and services (SBGs)*. While much work has been done to form theoretical and empirical basis for global poverty alleviation and the development of under-resourced communities, we feel that far too little attention has been dedicated to the study of how information and incentives drive learning, and how, in turn, the learning process deepens a populations commitment to adopt and consistently use SBGs.

The negative consequences of this lack of attention paid to adoption in mechanism and incentive/contract design is easily seen in the failure of costly interventions domestically and globally, where almost all of the effort is placed into mechanism design and automation practices (using digital technologies), and far too little effort is placed in understanding consumer behavior and psychology, sociological underpinnings of the intervened society, and classical economic thought and processes in the intervened locale, such that the procurement and *continued use* of the marketed good or service, shortly after the initial intervention period, becomes negligent, causing an overall failure of the intervention procedure, and a devastating economic and psychological loss to the spirit of altruistic poverty alleviation programs.

Our focus on incentive and contract design hinges on the importance of decision-makers having access to reliable preference sets of the various stakeholders. We call for and emphasize a renewed focus on listening to the reason and rationality of populations targeted in our intervention schemes, rather than prescribing (basic) rational behavior to simulated agents in our mathematical intervention framework.

Organization

In this chapter, we will introduce philosophical beginnings that formalize our goals and challenges for Optimal Intervention Theory (OIT). In the following chapter, we provide the most rigorous treatment of the models provided under OIT. We encourage the reader to refer to Appendix A (glossary) for a list of terms, before reading this and the following chapter.

4.1 The Problem

From a socioeconomic viewpoint, our main issue in sustainable global development theory for permanent, worldwide poverty alleviation is the fact that "no one wants to pay for positive externalities", or simply, that in "free-market" societies, producers are concerned only for the impact of systems (transportation, infrastructure, customs, software access, etc.) as they impact producers, and the same is said of consumers. Social benefit is a positive externality, and long-term dynamic contracting for the adoption of socially beneficial goods is specifically focused on maximizing the production of positive externalities. The "tragedy of the commons" that results from this fact has led Western (and other) governments to tax its citizens as a whole in order to fund things like University research, public education, public healthcare, government pension provision to combat elderly poverty (Social Security), and many other goods and services.

We clearly see, throughout the world, that such systems that aim to provide benefits beyond the producers and consumers (i.e., positive externalities) need large public funding concessions, or else will provide suboptimal service. Thus, we view a public-private partnership under the guise of the Principal Agent Model as the primary funding scheme in Heavily Indebted Poor Countries (HIPC) countries, especially [26]. We posit a global development theory where our objective is to provide the environmental characteristics necessary for the global citizenry to live *healthy and active lifestyles*, so that they can solve their own society's most pressing problems, which is the only sustainable practice for global development the author finds feasible in the long-run (vs. long-term dependence on foreign aid, tourism, and the like).

A Catalogue of Development Economics/Engineering Problems

Here, we provide an non-exhaustive list of current problems in development engineering and economics that lend themselves to system science and engineering based solutions, in partnership with experts from legal studies and business/finance. These disciplines (engineering, law, finance) must be interlinked in any sustainable development plan, to ensure the plausibility and integrity of proposed engineering solutions, the legitimacy of designed mechanisms, laws, and customs, and the liquid availability of assets necessary for development (as well as a low-cost, fixed rate funding scheme through public-private partnership).

1. Supply Chain Logistics
2. Trading infrastructure: Ports (air and sea), transportation vehicles (ships, trains, trucks) and networks, customs personnel, software, and practices, contracts and bargaining agreements, resilience plans and supply redundancies
3. Transportation networks, especially efficient last-mile delivery in "specially-constrained" environments
4. Interface of technology and development

   Fintech

   Biotech

Food, water, and energy technologies

Online optimization and personalization predictions

5. Well-privatized consumer identification schemes (for resource provision)
6. Software for trade and development
7. Education and training for trade officers, merchants, and citizens
8. Effective, localized consumer behavioral understanding and models
9. Private-Public funding schemes
10. Private-Public research & development schemes
11. Academic-based evaluation schemes

    especially those used to pivot/update interventions
12. Government management and peace relations
13. Local, regional, national, and transcontinental organization of trade unions and low-tariff contracts (for a specified development period, a "golden age" of global development)

Closely related to the areas for engineers, economists, philanthropists, investors, and other interested parties given above are the *statistics* for the end-users (consumers in developing areas) that detail progress in local and regional development. In C. West Churchman's book, *Challenge to Reason*, he provides a very succinct explanation that elucidates the difficulty of measuring human progress objectively [24]. In order to avoid contention and resources dedicated towards defining development goals, we feel that overall development goals, and thus statistics of progress, should be measured according to the enumeration provided by the UN's Sustainable Development Goals, and are thus referenced in this work and defined in our glossary [91]. Whereas Maslow's hierarchy still identifies relevant physiological needs (food, shelter, water), in this current globalized society, *transportation and energy access* must now be included as necessary to promote healthy and active lifestyles, and other considerations may be necessary.

4.2 Economic Theory and Contributions

In OIT, our sole purpose is to improve the utility of the agent, in some "markets" the principal cares about (i.e. desires a high utility for the agents in that market), over time, through the use of adaptive processes (i.e. exchanges between the agent and principal). To this end, this section introduces work from Roy Murphy on adaptive economic systems [72], R.H. Coase on transactions cost [25], and C. West Churchman and others on Game Theory [24].

We take the Coase based view of markets with "transactions costs" being the fundamental economic driver of economic processes, rather than the Robbins' view of allocation of scarce resources for maximizing utility and, in commonplace now, point to the 2019 Nobel Prize in Economics as reason to study interventions under this framework [25]. We do, however,

take the bidirectional hypothesis that the rate of generation of information in the the economic system controls the rate of decision making (i.e. time epochs where the contract is "renegotiated"), and thus indirectly the rate of growth of the system's resources (utility). This gives us direction for the formulation of the dynamic intervention process, and those things which effect the rate of change of the resource (or system utility) of interest.

Murphy and Adaptive Processes in Economic Systems

This section introduces work by Murphy, and defines what we describe as *adaptive intervention processes* [72]. In the preface of Murphy's work, he points out four issues with economic theory at that time: (1) Partial information is de facto where full information is de jure i.e. that most models assume full information. Full information is useful, however, for the study and construction of the basic structures of human and physical systems, in which we wish to show the power of full information. However for policy purposes, this in no way can be a rational casting of the problem. (2) Adaptive economic processes have not been studied in generality. (3) The theory of communication and information exchange must be treated more thoroughly, with transactions cost. (4) The analogies of an economic market under stochastic exchange and other scientific models with a stochastic number of events, each with associated payoff/utility, has not been properly exploited for modeling and analysis contributions.

Game Theory

C. West Churchman defines game theory as "a theory that tries to describe rational conduct in the context of conflict" [24]. Furthermore, he states, "It sets forth 'rational' rules governing "fair play" and tries to discover optimal strategies that can be followed by an individual when he is facing his opponent in a situation governed by certain recognizable rules." He believes, as do we, that the rules set forth here are still ineffective at helping management scientists make rational decisions (or policies).

According to these beliefs, we posit that operations engineering thinking, mixed with Game Theory, is the right academic basis for envisioning, designing, building, activating, and evaluating societal scale interventions. These are the "big scientific problems of our day". Utilizing data and engineering, mathematical, and statistical process to design intervention pathways for the adoption of social or population-based beneficial goods. There is an especial interest in designing systems that are *tractable*, meaning their impact lasts far past the intervention period, as measured by the indicators used to determine progress during the intervention cycle.

Churchman writes "In other words, the axioms of rational behavior provide the way of processing the basic information for rational plans, provided that information concerning preferences is available" [24]. Therefore, our focus on incentive and contract design hinges on the importance of decision-makers having access to reliable preference sets of the various stakeholders. We call for and emphasize a renewed focus on listening to the reason and rationality of populations targeted in our intervention schemes, rather than prescribing (basic) rational behavior to simulated agents in our mathematical intervention framework.

We make one final note in this section on the importance of population "typing", a clustering procedure, and its effect on intervention procedures. Now many of these interventions,

taking food systems as an example, can only be adopted in marginalized communities, such as those qualifying for SNAP or WIC benefits, due to our ability to influence them directly. However, for both the long-term entrenchment of the intervention goals and consistency (in the adaptive sense) over the intervention, socially beneficial goods and goals should be adopted for large segments of the population. Again, returning to food, although only x% of Americans qualify for SNAP, y% of Americans are overweight or obese. Clearly, then, we need to consider spillover effects, both negative and positive, over the lifetime of our intervention, and we may want to take the successes gained from adoption of those we can directly influence, and expand the program, through neighborhood-based models (e.g. community colleges, public broadcasting), to include the surrounding populations, and mainstream populations buffered from the margins of society. This idea has both moral (intervening on the most intervened, causing more instability) and practical applications (i.e. tractability of intervention procedure, herd immunity type arguments for overeating, lack of exercise, irregular sleep patterns, etc.

Transaction Costs and the Social Economy

Nobel Prize winning economist R.H. Coase is most widely recognized for his view of economics through the lens of transaction costs, and in his seminal work " The Firm, The Market, and The Law", Coase introduces the idea of transaction through both the perspective of the firm and market, stating of the market "Markets are institutions that exist to facilitate exchange, that is, they exist in order to reduce the costs of carrying out exchange transactions"; and of the firm writes, "The limit to the size of the firm is set where its costs of organizing a transaction become equal to the cost of carrying it out through the market" [25]. Now, these precepts of transaction costs, when applied to the social economy, gives us a deeper understanding of the work needed for successful social, political, and economic (SPE) system realignment in this lifetime.

Since transaction costs of many types (safety, infrastructure, education status, language barriers, etc.) are very high in the areas of our country, and in the developing world, where we will seek to intervene, we must invest in model programs with the understanding that they will have most of their resources absorbed into transaction costs. This phase of intervention is called *capacity-building*. However, these resources need not be wasted if the object of their use, the betterment of a certain locale for quality of life indicators, is utilized well in subsequent interventions in similar locales. With a large enough pool of model programs and interventions, we can utilize this information in dissimilar locales, once we are able to measure the specific interventions and their benefits in our model programs. This, again, brings in the ideas of statistical learning theory, to assure the methods and procedures added to the docket of best practices has been rigorously proven, and allowed the sensitivity to mold itself into a different locale, region, or culture. This is the idea of "measuring" or "managing" intervention dosage.

Social Welfare Functions (SWF's) and the Ramsey-Cass-Koopmans (RCK) Model for Economic Growth

RCK Model Applications

The Ramsey-Cass-Koopmans model has been used several times in the literature to model long term economic growth. In a study investigating long run output growth of multiple countries, the RCK model was used in the empirical analysis. Results reveal that the countries' growth levels do not converge using the RCK model.[1]

One of the assumptions of the RCK model is the homogeneity of the customers. In 2000, tools to study the evolution of the distribution of consumptions, assets and incomes in a market with heterogeneous customers were developed and applied as an extension to the standard RCK model.[22]

The RCK model has been used to investigate the positive correlation between saving and growth across different countries. In this study, precautionary saving motives are introduced into the basic Ramsey-Cass-Koopmans growth model in a manner that keep the model tractable. In addition, the author argues that the adding precautionary saving motives to the standard growth model can imply that increase in growth can cause increase in saving [54]. For this portion of the research work, we concern ourselves with firms (or groups of agents) characterized as nonprofit organizations in the American society, and have sought to quantify their decision-making processes. Most notably, we seek the formation of utility functions which address a nonprofit firm's tendency to balance these three business goals:

1. *Own Provision*: Firms maximize their own provision of goods $\min(Q_n, D_n)$, where Q_n is the amount of the good supplied by the firm and D_n is end consumer demand for their good [87] [2].

2. *Market Surplus*: Firms maximize end consumer market surplus CS_m [56] [104].

3. *Net Revenue*: Firms receive utility from net revenue $R_n - C_n(Q_n)$. This monetary return is primarily passed to firm members through higher salaries or benefits [79] [30].

Here, we seek to present tailored examples of the use of the Ramsey-Cass-Koopmans model, known classically as the Ramsey model of economic growth. The model's main formula is given as:

$$\max_c U_0 = \int_0^\infty e^{-(\rho-n)t} u(c) dt \qquad \text{(RCK)}$$

$$\text{subject to} \quad c = f(k) - (n+\delta)k - \dot{k}$$

where the constraint represents the evolution of capital accumulation, and our value function is some social planner's social welfare function. Capital accumulation (human and physical) in the world of OIT is paramount, and closely related to the idea of capacity building and intervention sustainability. Any and all intervention efforts must seek to raise, maintain, and permanently entrench human capital systems (i.e. education and training based-systems, as well as retention and recruitment of personnel) and physical capital systems (transportation networks and vehicles, customs and customs software, structures and architectural design, energy and power systems, water source management and discovery, etc.).

Therefore, in this research work, we seek to illuminate the potential value of Social welfare functions, with a focus on the Ramsey-Cass-Koopmans model and its applications. The model discounts social utility over time, with respect to a firm's (agents) ability to accumulate capital while fulfilling the firm's (principals) specific utility function (social and economic motives). The idea of a firm must be gathered from context here (i.e. when it describes the agents goals of the PAM, and the principals). We explore different solution methods that have been implemented to solve this problem. In addition, we present tailored examples of the use of the Ramsey-Cass-Koopmans model, known classically as the Ramsey model of economic growth, in the context of Non-profit organizations and climate change economics. We discuss the formulation of these two problems, with particular emphasis on construction considerations for the utility function critical in the RCK model.

This model provides a method to consider trade-offs between altruistic and capitalistic business practices, which may be combined in a manner beneficial to equilibrium-based markets. Researchers believe that the necessity to accumulate capital to fulfill any business goal can be quantified, and therefore, optimized for business sustainability and growth. The author intends to make the following contributions to OIT using this theory, which encompass both technical considerations and social welfare notions:

1. Catalog some existing solution methods to the RCK model
2. Present Models for tailored examples using RCK in the context of Non-profits and discuss potential solution implications.

A literature review of the Ramsey-Cass-Koopmans model and its applications to date can be given by the author for any interested parties.

Ramsey-Cass-Koopmans (RCK) model

The standard neoclassical models of economic growth are those associated with Ramsey, Cass and Koopmans, in which growth is a function of saving, investment and capital accumulation. The Ramsey-Cass-Koopmans (RCK) model aims only at explaining long-run economic growth rather than business cycle fluctuations, and does not include any sources of disturbances like market imperfections, heterogeneity among households, or exogenous shocks. Originally Ramsey set out the model as a central planner's problem of maximizing levels of consumption over successive generations. Only later was a model adopted by Cass and Koopmans as a description of a decentralized dynamic economy. Subsequent researchers extended the model, allowing for government-purchases shocks, variations in employment, and other sources of disturbances, which is known as real business cycle theory.

RCK Model Formulation

The Ramsey-Cass-Koopmans model is an extension of the Solow growth model whereby the new feature is that saving rate is not exogenously given. The objective of the RCK model is for firms is to maximize profits while maximizing individual/household utility [31]. This model has several underlying assumption, stated below:

- Capital is endogenous while knowledge and labor are exogenous

- Capital and output are the same commodity. Thus, capital can be consumed
- No depreciation (of capital)
- Households earn profits as if they owned the firms
- Saving and consumption are endogenous.

RCK Model Solution Methods

Solving this problem, for instance by converting it into a Hamiltonian function, yields a non-linear differential equation that describes the optimal evolution of consumption,

$$\dot{c} = -\frac{u_c(c)}{c \cdot u_{cc}(c)}[f_k(k) - \delta - \rho] \cdot c$$

which is known as the Keynes-Ramsey rule.

One of the methods of solving the RCK model is by simulation. The core equation in the RCK model is the capital and wealth accumulation differential equation. The RCK model is simulated by solving the set of two coupled ODE's. A non-negativity constraint is imposed on the solutions to the two equations. A trajectory of the solution path, showing the time evolution of capital and consumption is created. Steady states are obtained by setting differential equations to zero. [18]

Application of RCK model to Nonprofits

In this section, we discuss applications of the RCK model as Social Welfare Functions (SWF) for specific applications in the context of non-profit organizations. With reference to the infinite period model in [30], we have a discounted utility function of the form

$$V(X_{nt}, S_t|\rho_n, \rho_{-n}) = \mathbb{E}_{\rho_n,\rho_{-n}}\big[\sum_{k=t}^{\tau_i} \beta^{k-t} U(X_{nt}, S_t) + \sum_{t=\tau_i}^{\infty} \beta^{k-t} Y_{nt}|X_{nt} = X, S_t = S\big]$$

where τ_i is a random stopping time based on some nonprofits strategy profile, denoting the period in which they leave the market (for outside option Y_{nt}) Then we seek to maximize this value function with regards to a Markov strategy (for each firm) $\rho_n : S \Rightarrow A_n$, for a set of available options $\mathbf{A}_{nt}$ for firm "n" at time t, and a market state $S_t \in S$. The assumed constraint in the original model is written as

$$\mathbb{E}_{\rho_n,\rho_{-n}}\big[\sum_{k=t}^{\infty} \beta^{k-t} C_{nt}(X_{nk}, S_k)\big] \leq \omega_{nt} + \mathbb{E}_{\rho_n,\rho_{-n}}\big[\sum_{k=t}^{\infty} \beta^{k-t} R_{nt}(X_{nk} S_k)\big]$$

where ω_{nt} is the assets of the firm in period t and $\beta = \frac{1}{1+r}$, where r is a constant real interest rate. As this value function takes the form of a utility function, to be maximized, its application to the RCK model is apparent. Hence, we move to form constraints (time-varying) that align with the RCK model's, and to describe and construct some considerations on forming a concrete utility function for community-based nonprofits (excluding large, NGO-like organizations such as large art studios, missionary organizations, etc. whose budgets exceed most corporations.)

Design of the Utility Function

As the notion of creating a utility function dependent on some universal set of firm characteristics is futile, we attempt to make considerations for "most" firms that fit within our frame of community-based nonprofits. Here are a list of firm characteristics our model wishes to quantify for social impact:

1. Service to the surrounding community

   Metric created to analyze the demand for the good in the Nonprofits immediate vicinity, and how much of this demand is met. This notion yields a beneficial social surplus for nonprofits located in areas where the demand for their good or service is high. *Mathematical Representation*: This variable can be represented as a normalized ratio describing the amount of local demand met by the organization, with regards to the organization's human and physical capital. The score can be thought of as an "Own Provision" scheme where we seek to $\min(Q_n, D_n)$ where D_n is some forecasted demand and Q_n is the proportion fulfilled by Nonprofit firm i.

2. Service to the (regional and) national community

   Metric created to analyze the demand for the good outside the Nonprofits immediate vicinity, and how much of this demand is met, with considerations for international impact as well. *Mathematical Representation*: This variable can be represented in the same manner as the first descriptor, with special attention paid to the *normalization* of this ratio based on business size and (general field) scope.

3. Policy Impact

   Local, Regional, and National measures, likely in a hierarchal system as listed, with National impact being regarded as the most important to the measure. *Mathematical Representation*: This predictor variable will be a threshold-based Bernoulli that measures whether or not a Nonprofit's leaders and methodologies affect national conversation and/or legislation in the firm's target area.

4. Donor consistency

   We consider a measure of a philanthropic gift's resonance with the donating group, after the completion of the contract work. This can be measured on a standard Likert scale. *Mathematical Representation*: Likert Scale based on the survey of past Donors, with answers averaged.

5. Donor versatility

   We consider a measure of versatility for donations with regards to the field or target area of a philanthropic organization. This can be (initially) modeled as an indicator variable with some positive weight (adding value to the company). *Mathematical Representation*: This variable may be one of the hardest to quantify, but should consider both the firm's past and current donor versatility, as well as the areas in which they may have some future impact, based on current business assets. We suggest a Bernoulli indicator based on this criterion at first, and the development of more accurate measures when the model is created.

These characteristics describe a Nonprofits characteristic vector $\mathbf{x}_{nt}$ at time epoch t. Given available data in the area, we may wish to construct a multiple regression model with linear function $\sum_{i=1}^{5} \beta_{nti}\mathbf{x}_{nti}$ For now we assume the proper tuning of these weights (which would be industry specific), such that

$$U(X_{nt}, S_t) = \beta_{nt}\mathbf{x}_{nt}. \tag{4.1}$$

In this section, we hoped to motivate consideration of Optimization techniques that can be developed to accurately qualify social welfare in the contexts of market equilibrium-based economics. The explored topics represent only two of many social ills for which a philanthropic based market could help realign imbalances. Furthermore, we conduct research in the era of Big Data, which will allow the active training of our models using real-world data, and, when necessary, could find it useful to help generate some the data with the existing firms in this sector.

As a first step to the study of this area rigorously, in the next chapter introduce notation from several fields to build a coherent notational base for adaptive intervention processes. The main tributary field is dynamic programming (DP), for models and notation. Other fields of optimization theory, especially stochastic (i.e. with randomness), such as optimal control and queuing theory, as well as thee economic models presented here, will be consulted for the task.

Chapter 5

A Mathematical Framework for OIT

Following the introduction to models and modelling concepts from the social sciences in the previous chapter, we will fully develop the mathematical elements introduced, and provide results from outside literature that will help us form a holistic theory. We provide narrative where appropriate to connect these results and formulations to the design and analysis of societal scale interventions.

In this chapter, we present the main model of OIT, which is an intervention model whose output is a dynamic optimal contract of interventions between a principal and agent under a time-indexed version of the quartic knapsack problem. Recall that the main thrust of adaptive processes is that decision makers (or processes) increase their knowledge by the cumulative experience of "doing while learning". Therefore the models we have seen up until this point will now be revisited. In particular, sample average approximation results for stochastic problems involving ℓ_1 constraints will be a part of the OIT model, and our sample bounds in that chapter will be relevant for estimating the principal's investments over the intervention period, as well as learning the principal and agent's true utility functions over time. The tensor completion problem represents the work we can do with data structures which hold our realized random variables' value over time, again helping us to predict the future value random variables through completion, and also gives us insights through the direct use of factorization techniques, similar to singular value decomposition in the matrix world.

5.1 Mathematical and Statistical Formulations and Analysis

This section provides formal theory to *bridge the gap between classic, expert-based iterative methods for global poverty alleviation and international development, and new data-driven approaches that collect and analyze large amounts of data to form prediction and classification models that increase intervention efficiency.* In particular, classic social science and business-like approaches to the alleviation of global poverty and its associated development issues may be described as a "theory of experts", and the nature of NSF and other like organizations' calls for proposals where multidisciplinary teams use data-driven approaches to solve national-scale problems are an example of our nations' experts' commitment to a new academic and research

regime. "In-between" the meetings and offerings of these experts in multidisciplinary teams, experts in data science, machine learning, algorithm development, cybersecurity, and other mathematical and statistical approaches to data utilization, combined with the large amounts of data and computing power we now have access to, will help to "fit" the content offered by these experts to the current realities for the populations we intend to impact.

5.2 Optimization Models and their Interpretations and Uses for OIT

The Principal-Agent Model

Following the structure of the Principal-Agent Model (PAM), where a principal contracts with an agent to do some task the principal cannot see (moral hazard) and/or cannot estimate effort properly (hidden information/adverse selection), *in OIT, our sole purpose is to improve the utility of the agent, in some "markets" the principal cares about* (i.e., desires a high utility for the agents in that market), over time, through the use of adaptive processes (i.e., learning-based exchanges between the agent and principal). Examples can be seen with governments and economic markets, such as in agriculture and financial technology (fintech) markets, especially in the developing World [3]. Perhaps a better example is the government contracting of external parties for defense (e.g., Raytheon, GE, Locheed Martin, etc.). Then, in any contractual agreement between principal and agent in this view, must begin with answering the three basic questions for adaptive behavior [72]. Importantly, we view our PAM as a two-way contract, in which *the principal must use the adaptive process to improve its behavior and actions towards the agents as well.*

1. Under what conditions does the adaptive process always improve the behavior of the decision makers (principals and agents)?

2. What controls the rate at which the expected improvements in behavior occur (esp. for agents)?

3. Can the adaptive process explain (e.g., by Principal Component Analysis) the diversity of observed behavior of supposed rational decision makers without appeal to the existence of (unknowable) individual utility functions?

In particular, we will focus on societal scale interventions for adoption of socially beneficial goods (SBG's), services and actions. To begin the theoretical study of human behavior under this theme, causal relationships are not necessary, and correlative effects are important, as our primary concern is the efficiency of the agent. With "ensemble" intervention methods (many different models combined), it will often be difficult to know which factors were most important, until we have generated lots of (unique) data. The Quadratic Knapsack Problem (QKP) does give some structure to measuring *synergistic* effects; that is, the impact that individual interventions may have in combination, which can be negative as well (i.e., Negative feedback loops have negative synergy) [80].

We will also use an ensemble method in measuring the effectiveness of intervention strategies; whereas most models focus on only one intervention goals (access-, resource-, education-, or network-based), our models will maximize a combination of the overall, average, and maximum socially beneficial character-actions and character-states. Then, following Paulson's example for the bi-level program for optimal interventions [60], the upper level of our optimization models will measure some statistic related to the overall adoption of SBG's in our intervention context, and the lower level will consist of a consumer behavioral model which combines known qualitative and quantitative information on the intervention context and agents, in order to move them along an optimal path towards adoption. Here is the main contribution of our work and similar, nascent work. To expel the academic-centered "rational actor" model and institute a consumer behavioral model that uses the latest information provided from psychology and other social sciences to accurately reflect the psyche (or strategic behavior) of agents and *types of agents* in new intervention contexts [8].

The mathematical beginnings for the PAM can be found in the incentive theory work by Laffont [55]. The book gives a mathematical basis for considering contracting over time through the PAM, including analysis of systems with adverse selection (privately held information) and moral hazard (inability to observe the agent's work). What is unique and relevant in our model is that *we consider interventions where our work is to bring the agent from the starting characteristic state to an agreed-upon ending characteristic state*, over the contract period, rather than the typical production-based principal agent relationship. Therefore, our optimal contracting policy cannot be based upon acute actions from the principal to the agent, but instead we require that the agent act for itself for the overwhelming majority of the contracting period, to help itself reach the agreed upon pre-intervention "success state", over the contracting period.

This type of intervention style is especially relevant to our main application area, diet-based health disparity alleviation. In these cases, such as for diabetic patients, the agent (patient) must make the majority of its actions (what to eat, exercise, sleep patterns, etc.) on its own, unobserved by the agent. Then, our job in *contract design* is to set *incentives* in such a way that the agents have motivation to act, often, in their best interest (e.g., $\sim 80\%$ of actions are positive). A combination of education-, resource-, access-, and network-based (or group/type-based) interventions can be offered from the principal to achieve this goal. We introduce concepts that rely heavily on the agents' type and past actions, which, under the guise of operations research, is referred to as the agent's historical information vector, or *filtration* (or sigma algebra).

We will provide a complete treatment of the use of the PAM later, but will provide basic formulas for the principal and agents' utility functions (i.e., incentive structures), and introduce terms detailing the action space of the principals and agent in a dynamic, repeated-game setting. Here, the agent takes two types: Efficient ($\bar{\theta}$) and inefficient ($\underline{\theta}$), and we assume an agent is efficient or inefficient with probabilities v and $1-v$ respectively. This can be easily generalized (and will be) to include more types. In OIT, we study socioeconomic problems where both the principal and agent can influence (i) *their individual effort levels*, and thus payoffs, according to the a priori, agreed upon compensation structure, as well as (ii) *their environments*, by "economic" methods such as system investments, and "social" methods such as updating their utility functions, for example, to provide more reward (or value) for producing social good/positive externalities.

According to Laffont's work, the principal's general payoff (optimization) function is with respect to a "menu of contracts", and takes the form:

$$\begin{aligned} \max_{\{(\bar{t},\bar{q});(\underline{t},\underline{q})\}} \; & v(S(\underline{q}) - \underline{t}) + (1-v)(S(\bar{q}) - \bar{t}) \\ \text{subject to}: \quad & \underline{t} - \underline{\theta q} \geq \bar{t} - \underline{\theta}\bar{q} \\ & \bar{t} - \bar{\theta}\bar{q} \geq \underline{t} - \bar{\theta}\underline{q} \\ & \underline{t} - \underline{\theta q} \geq 0 \\ & \bar{t} - \bar{\theta}\bar{q} \geq 0 \end{aligned} \tag{5.1}$$

where $\bar{t}$ and $\underline{t}$ are the transfers (payoffs) for an efficient and inefficient agent in a single period, respectively, and $\bar{q}$ and $\underline{q}$ are the quantities of goods (or service level for service products) produced in that contracting period. Constraints 1 and 2 are called incentive compatibility constraints, and constraints 3 and 4 are called participation constraints. In some (many) contexts, one or both of these sets of constraints are unnecessary, but they're meaning is important to interpret. Definition 2.2 of the text says that a menu of contracts is incentive feasible if it satisfies both incentive and participation constraints, and that these constraints *fully characterize* the set of incentive feasible menus of contracts.

The agent's general payoff (objective) function is utility maximizing, and takes the form.

$$\begin{aligned} \max_{\{(\bar{t},\bar{q});(\underline{t},\underline{q})\}} \; & U = t - C(q,\theta) \\ \text{subject to:} \quad & C(q,\underline{\theta}) = \underline{\theta} q + F \\ & C(q,\bar{\theta}) = \bar{\theta} q + F \\ \underline{U} = \underline{t} - C(\underline{q},\underline{\theta}) & \geq \bar{t} - C(\bar{q},\underline{\theta}) \\ \bar{U} = \bar{t} - C(\bar{q},\bar{\theta}) & \geq \underline{t} - C(\underline{q},\bar{\theta}) \\ \underline{U} = \underline{t} - C(\underline{q},\underline{\theta}) & \geq 0 \\ \bar{U} = \bar{t} - C(\bar{q},\bar{\theta}) & \geq 0 \end{aligned} \tag{5.2}$$

Here the first two constraints define the agent's cost functions, given its respective type, and the last four constraints are again incentive (compatibility) and participation constraints. The book and most papers on the PAM begin to discuss the optimal contract after these incentive and modeling notions are understood and represented mathematically.

Then, the general PAM has an objective where the principal maximizes its payoff subject to eliciting a minimum effort level from the agent, for which they compensate the agent. This notion is called the "The Rent Extraction-Efficiency Trade-Off", and is the title of chapter 2 of Laffont's work, previously mentioned. We will view the dynamics between an agent's observed actions, and the principals response to either continue in contract, renegotiate, retire the agent, replace the agent, or promote the agent, and similarly the agent's decision to continue in contract, or quit (abandon the system), as decisions made in an infinitely repeating, two-player game. Chapter 8 of Laffont's work gives us base models for this regime [55]. It is simple to make the reward structure zero sum, or use other game theoretic notions that lend themselves to useful and interpretable analysis.

Dynamic Programming (DP) and Optimal Control

One of the most powerful tools from statistical learning theory for OIT is Dynamic Programming (DP) and optimal control, and its subsequent derivations, such as Model Predictive Control (esp. learning-based). The DP algorithm solves problems in *stages*, and calls for the "balance", or trade-off, of the current stage's decision with an approximation function for all future periods. In other words, the algorithm calls us to be optimal, in a very specific sense, over all possible decisions in our current state, and to *approximate* the cost of the remaining stages, as to avoid a decision in the current stage that causes us to incur suboptimal cost in the future (due to a deviation from the optimal "path" of decisions). The basis of this work has been consolidated into the well-known work by Bertsekas, which is printed in two volumes [14]. First, we discuss aspects of the DP algorithm and how they relate to OIT. Dynamic programs can be completely characterized by (1) an underlying discrete-time dynamic system (can be continuous, and is discretized using the notion of entropy time), (2) a cost function $[g_k(x_k, u_k, w_k)]$ that is additive over time, related to the current system. The system has the form

$$x_{k+1} = f(x_k, u_k, w_k), \quad k = 0, 1, \dots N-1$$

where k indexes time, x_k is the state of the system and summarizes past information that is relevant to future optimization, u_k is the "control" or decision variable to be selected at time k, w_k is a random parameter (also called disturbance or noise depending on the context), N is the "horizon" or number of times control is applied, and f_k is a function that describes they system and in particular the mechanism by which the state is updated. This is all taken from the introduction in chapter 1 of volume one of Bertsekas' work on Dynamic Programming and Optimal Control [14]. Next we provide interpretations for OIT.

System State, x_k

For the purposes of OIT, our system state, x_k, describes the summation of the information we have concerning a population for which we desire altruistic, benevolent intervention. The theoretical underpinnings of *data generation* are of import here. We describe the importance of a few categories of measurement that apply to human systems we hope to optimize. Communities of people can be "typed" (or classified) based on these measurements, through simple procedures such as k-means or k-median clustering [49], or more sophisticated data clustering tools/models [44]. All of our interventions fall into four categories, namely: (1) Access, (2) Resource, (3) Education, (4) Network (Group-based). Herein is the first reference to *integer programming*. Many of the states above can be crudely measured as binary variables describing the presence or non-presence of specific access, resource, or educational availability to the individuals and communities involved. Using the example of food nutritional interventions, "access" may refer to the proximity of grocery stores and restaurants (food deserts [12]), or to cookware, energy, and water availability for needed for cooking and food storage. Then, the simplest consumer behavioral descriptions exist within the convex hull of 0-1 polytopes, where vertices represent known (a priori) resources that citizens may or may not have available to them. The fourth list item is also interesting, as network analysis often lends itself to integer, linear, and mixed integer linear approaches.

Something to be studied in this field of information transfer and availability, and its relation to learning and adoption of SBGs, is the effect of network (community) influencers on the intervened population, which can be gauged classically by weighting nodes representing individuals by the cardinality of their connections (i.e., how many people they are connected to), or the nature of their connection. For example, and again looking at food nutritional interventions, the grocery purchaser and preparer of foods in a household is the most important influencer for a family's nutrition, and our models (and past interventions) would suggest, by network analysis, that this is where our *controls*, or interventions, should be focused in this context. Extrapolations can be made for different intervention environments and tasks.

In constructing behavioral models in new intervention contexts, we will be concerned with people-based system states, which we have and will refer to as "characteristic" or "character" states. These notions lead us to population "typing", or group classification. Some of these classically exist and are used in policy and intervention practice, most notably things like race, socioeconomic status (SES), and other demographic notions, as well as marital status and number of dependents (important for time and resource availability), and participation in affinity and belong groups. There is also subcategories related to people that relate to the four overall categories given above. For education, for example, this includes formal and informal education (myths, cultural barriers, etc.). From our studies in international interventions, many have concluded that "Most people, as it relates to interventions, are either not ready, unsure, or ready (to adopt)". Each of these states requires different methods for state improvement.

The DP algorithm is typically written, for N stages, as:

$$\begin{aligned} J_N(x_N) &= g_N(x_N) \\ J_k(x_k) &= \min_{u_k \in U_k(x_k)} \mathbb{E}_{w_k} \left\{ g_k(x_k, u_k, w_k) + J_{k+1}(f_k(x_k, u_k, w_k)) \right\} \\ & \qquad\qquad k = 0, 1, \ldots, N-1 \end{aligned} \tag{DP}$$

The interpretation for OIT is this: There is a fixed end state (more specifically, character-state), x_N, which we want agents and groups of agents to reach in some specified time period. The "stage cost" for this end stage is fixed as $g_N(x_N)$. Similarly, stage 0 consists of given initial conditions, which will describe the consumer behavioral models at the beginnig of our intervention period. Then, the DP algorithm proceeds at each stage "k" to minimize the "expected" value of the current stage costs (say, an intervention subinterval of 1 month) with a prediction/estimation of future costs (the rest of the life of the intervention scheme). Overall, with a well-defined end-goal in mind, say for diet-based health disparities, an end-goal representing consistent healthy eating, exercise routines, and sleep/rest patterns, we want to move as much as we can in one stage towards this optimal end-goal, while keeping in mind that we can take actions in future stages to recenter our agents on the optimal pathway (if they have deviated), or to further encourage optimal behavior in the next period (if they fulfilled all goals for that subinterval). The following proposition is taken from the text given above, and formally explains the use of the DP algorithm.

Proposition 17 (Proposition 1.3.1 of [14])**.** *For every initial state x_0, the optimal cost $J^*(x_0)$ of the basic problem is equal to $J_0(x_0)$, given by the last step of the above algorithm, which proceeds backwards in time from period $N-1$ to period 0.*

The expectation in the second equation of (DP) *is taken with respect to the probability distribution of w_k, which depends on x_k and u_k. Furthermore, if $u_k^* = \mu_k^*(x_k)$ minimizes the right hand side of the second equation above for each k and x_k, the policy $\{\mu_0^*, \ldots, \mu_{N-1}^*\}$ is optimal.*

This proposition and its use in "deterministic" settings (i.e., optimization problems where all information is known and unchanging from the outset of the problem) means that we can move "backwards in time" from our well-defined end goal, optimizing the actions at each stage, until we have reached stage zero, at which time the optimal backwards path can be inverted for a forward path. This is taught most often with small routing problems in DP classes, including the Traveling Salesperson Problem [45]. For stochastic problems (i.e., imbued with randomness, some of which can be estimated using probability theory and statistics), which are our interest, *as human behavior always introduces stochasticity*, we cannot determine optimal paths from the beginning of the problem. Perhaps the best known DP problems with stochasticity today are the problems solved by connected and automated vehicles (CAV's, self-driving cars), in which they predict obstacles and conditions that may affect their travel path, but still must attune in every stage (which is usually about 5 second intervals) to the *realization* of those predicted, stochastic elements [7]. Similarly, in creating dynamical systems that measure optimal interventions for OIT, we must predict random human behaviors to the best of our ability for current and future periods, and make adjustments.pivots as more behaviors are realized, which will be captured in our information-gathering procedures, over the lifetime of our interventions.

The Quadratic Knapsack Problem and Extensions to Quartic

In this section, we will first introduce the quadratic knapsack problem and discuss its improvement to the classic linear Knapsack problem, which is the introduction of *synergistic effects*, which are benefits (or consequences) gained from the combination of individual items in the knapsack, which will represent our intervention levers (bandit-arms) available to tthe principal for any given stage in the intervention process. The binary quadratic knapsack problem (QKP) was first introduced by Gallo et. al. Its formal definition is as follows: Assume that n items are available, where item j has a positive integer weight w_j. Additionally, we are given an $n \times n$ nonnegative integer matrix P $=\{p_{ij}\}$, where p_{jj} is the profit achieved if item j is selected and $p_{ij} + p_{ji}$ is a "synergy" profit achieved if both items i and j are selected for $i < j$. Given the indices $[n] := \{1, \ldots, n\}$ to denote the item set N, we introduce a binary variable x_j to indicate whether item j is selected. The QKP calls for selecting an item subset that maximizes the overall profit, while fulfilling the knapsack capacity (budget) constraint.

The quadratic knapsack problem (QKP) can be written as:

$$\begin{aligned} \max \quad & \sum_{i \in N} \sum_{j \in N} p_{ij} x_i x_j \\ \text{s.t.} \quad & \sum_{j \in N} \omega_j x_j \leq c \\ & x_j \in \{0, 1\}, j \in N \end{aligned} \tag{QKP}$$

Without loss of generality we assume that $\max_{j\in N}\omega_j \leq c < \sum_{j\in N}\omega_j$ and that the profit matrix is symmetric, i.e., $p_{ij} = p_{ji}$ for all $j > i$.

Note, some formulations write the objective as

$$\max \sum_{i=1}^{n} p_i x_i + \sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}} P_{ij} x_i x_j$$

to make the synergy profit more explicit; however, this is covered in our formulation for the construction of P as described. In particular, our formulation has

$$p_{ii} \cdot x_i^\top x = p_i x_i,$$

and we can scale synergy profits to $\hat{p}_{ij} = p_{ij}/2$ so that

$$p_{ij} x_i x_j = \hat{p}_{ij} x_i x_j + \hat{p}_{ji} x_j x_i$$

for any pairs x_i, x_j selected in the knapsack (including the "double count" of singleton items). We prefer (QKP) for its compactness. Much work has been done in both the realms of nonlinear programming and continuous relaxations for the 0-1 knapsack problem. For a course project, the author took a look at some of the leading contributors to the origin of solution methodologies detailed improvements made by several practitioners since the creation of these foundations. For those interested in Integer Programming and the problem in generality, the author can be contacted for provision of this full report.

Pisinger's survey [80] provides a comprehensive review on the development of QKP's upper bounds and their application in variable reduction, as well as some heuristics, approximation algorithms, valid inequalities, and branch and bound algorithms. The survey's main focus concerned the introduction of each upper bound to the field, and their incorporation into different solution techniques. The survey also provided a comparison of the efficiency of different upper bounds. The QKP is usually solved by deriving good upper bounds and using approximation techniques until some tolerance parameter is satisfied, i.e., we have an integer (binary) solution that is "close enough" to the upper bound. Techniques to reduce the complexity of deriving these upper bounds include relaxation from upper planes, linearization, reformulation, Lagrangian relaxation, Lagrangian decomposition, and semidefinite programming [80]. Since any feasible integer solution, gained from heuristics or exploitation of the problem structure, gives lower bounds, solution techniques focus on closing the gap between the upper and lower bounds [81].

This method will be extended in this next chapter to include four sets of intervention categories, which are access-based, resource-based, education-based, and network-based interventions. As mentioned in the beginning of this subsection, the power of updating the linear knapsack problem to higher polynomial dimensions is the inclusion of synergistic effects, which requires us to determine the relationship between the variables that represent our principal's actions in the intervention space. This method will be used to form the higher-level optimization problem (our principal's problem) and is naturally additive over time, lending itself to use as a cost function in an optimal control regime. Additionally, the idea of utility maximization subject to resource constraints is described by this model, and can be used in the construction of our lower-level (consumer behavioral) models.

Multiobjective (Simulation) Optimization

Based on the economic notion of solution *trade-offs*, there has formed, recently, a profound interest in developing engineering and financial optimization theory which produces *efficient sets* of solutions, rather than unique minima/maxima, to optimization problems. This interest has proliferated into the field in general, and has been consolidated into Multiobjective Optimization (MOO), in particular. The problem is generically written as:

$$\min_{\mathbf{x}\in\mathcal{X}} (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})) \tag{MOO}$$

MOO methods produce *efficiency fronts* that most often align with the notion of Pareto Optimality, although other definitions for efficiency in the context of optimization have been offered [105]. Similarly, *nondominated* points, or efficient points, are solutions to MOO problems for which there is no other (known) point that is at least as small in all objectives, and strictly smaller in at least one objective.

MOO methods are applied to problems which, unlike the convention of single objective optimization, consider multiple, generally conflicting objectives at once, and, from the desk of the problem-solver, offers multiple efficient solutions to the decision makers, each with their identifiable trade-off. The field, then, takes an important humanistic approach, in that it distinguishes the problem-solver from the decision-maker. Indeed, the role of the decision-maker and problem-solver are entwined under the guises of MOO, wheras single-objective optimization sees no divide here.

MOO is further subdivided into three categories in the annals of the current literature, based on the approach desired from the decision-maker. These three categories are based on the timing of the decision-maker's influence on the system, and can be described as:

- *No-preference*, in which any solution to the multiobjective formulation is acceptable;
- *a priori prefernces*, in which information provided to the problem-solver by the decision-maker is used in the formulation of the model, usually quantified by (weighted) scalarization techniques, which reduces the MOO problem into a single-objective optimization problem. However, more elaborate schemes than scalarization exist [cite].
- *a-posteriori preferences*, in which the decision maker reveals her preferences after the problem is solved to its finest, or most robust, precision. The decision-maker can then choose from the list of *non-dominated* solutions, i.e., those solutions which use all available resources efficiently, based on her preferences and the breadth of solutions presented by the problem-solver.

In a fuller work, we develop a method which addresses the third category above. An important note is that we address *misspecification* of the preference by specifying an *interval* around the decision-maker's preference. In this work, we introduce MOO for its use in our OIT model, specifically in determining future utilities for our principal and agent. We will take the a prioiri technique of scalarization as a basic introduction to MOO in our model, but have introduced the use of more sophisticated techniques here, including the computationally-heavy process of producing the entire efficiency front for a posteriori use by the decision maker.

5.3 A General Model for Human System Interventions

Following our discussion of the OR models that lend themselves to a global theory of intervention goal-setting and path-setting, under the guise of long-term dynamic contracting, we formally present our model and its implications for the field.

Notation

Let $\mathcal{I}_t$ be the set of all access-based ($\mathcal{A}$), resource-based ($\mathcal{R}$), education-based ($\mathcal{E}$), and network based ($\mathcal{N}$) interventions available at the beginning of time step t to form contract $C_t(\bar{t}_t, e_t, p_{C_t})$, with profit $p_{C_t} = \sum_t p_{ijk\ell}$ for the principal, which uses available strategies $(i,t) \in \mathcal{A}_t$, $(j,t) \in \mathcal{R}_t$, $(k,t) \in \mathcal{E}_t$, and $(\ell,t) \in \mathcal{N}_t$, and where transfer $\bar{t}_t$ is given (at the end of the period) to a contracting agent (or group of agents) for effort e_t. At each time step $t+1$, the principal changes the intervention environment by selecting a subset $\{i_{t+1}, j_{t+1}, k_{t+1}, \ell_{t+1}\}$ of interventions based on its valuation of the previous period's profits, which is a convex combination of a predictive consumer behavioral model,

$$\sum_{i \in A_t} \sum_{j \in R_t} \sum_{k \in E_t} \sum_{\ell \in N_t} u^*_{C_{t+1}}(x_{i,t}\, x_{j,t}\, x_{k,t}\, x_{\ell,t})$$

and the previous periods true profit function

$$\sum_{i \in A_t} \sum_{j \in R_t} \sum_{k \in E_t} \sum_{\ell \in N_t} p_{C_t}\,(\, x_{i,t}\, x_{j,t}\, x_{k,t}\, x_{\ell,t})$$

Alternatively, at each time step $t+1$, the agents change the intervention environment by selecting a subset $\{i_{t+1}, j_{t+1}, k_{t+1}, \ell_{t+1}\}$ of interventions based on a convex combination of its utility maximizing function

$$\sum_{i \in A_t} \sum_{j \in R_t} \sum_{k \in E_t} \sum_{\ell \in N_t} u^*_{C_{t+1}}(x_{i,t}\, x_{j,t}\, x_{k,t}\, x_{\ell,t})$$

and its belief of the evolution of the intervention environment, which includes changes its own utility function u_{t+1}, and its expected change in the principal's profit function $p_{C_{t+1}}$ at time $t+1$, and its expected change in the intervention environment (possible contracts $C_{t+1}(\bar{t}_{t+1}, e_{t+1}, p_{C_{t+1}})$

Our Model

Our model is a bilevel optimization model, whose higher-level objective is a time-indexed quartic knapsack problem (QuKP), that tracks individual and combinatorial investments/interventions into the access, resource, educational, and network (or group-based) variables representing the contracted population (i.e., agents) by the members of the *social economy* as principals, and whose lower-level objective is a consumer behavioral model with incentive and participation constraints from the PAM (enforced when necessary).

Inputs: Let $\mathcal{I}_t$ be the set of all access-based ($\mathcal{A}$), resource-based ($\mathcal{R}$), education-based ($\mathcal{E}$), and network based ($\mathcal{N}$) interventions available at the beginning of time step t to form contract

$C_t(\bar{t}_t, e_t, p_{C_t})$, with profit $p_{C_t} = \sum_t p_{ijk\ell}$ for the principal, which uses available strategies $(i,t) \in \mathcal{A}_t$, $(j,t) \in \mathcal{R}_t$, $(k,t) \in \mathcal{E}_t$, and $(\ell,t) \in \mathcal{N}_t$, and where transfer $\bar{t}$ is given (at the end of the period) to a contracting agent (or group of agents) for effort e_t.

Model: Quartic Knapsack Problem for a Dynamic (Adaptive) Principal-Agent Model

Output: Next period's contract (C_{t+1}) with selected interventions

$$(i, t+1) \in \mathcal{A}_{t+1}$$

$$(j, t+1) \in \mathcal{R}_{t+1}$$

$$(k, t+1) \in \mathcal{E}_{t+1}$$

$$(\ell, t+1) \in \mathcal{N}_{t+1}$$

The upper level objective is the principal's expected profit:

$$p^*_{t+1} = \max_{C_{t+1}} \sum_{i \in A_{t+1}} \sum_{j \in R_{t+1}} \sum_{k \in E_{t+1}} \sum_{\ell \in N_{t+1}} \mathbb{E}\left(p_{C_{t+1}} \ \left(x_{i,(t+1)}\, x_{j,(t+1)}\, x_{k,(t+1)}\, x_{\ell,(t+1)}\right) \mid p_t\right)$$

where p_t is

$$p_t = \lambda \sum_{i \in A_t} \sum_{j \in R_t} \sum_{k \in E_t} \sum_{\ell \in N_t} p_{C_t}\,(x_{i,t}\, x_{j,t}\, x_{k,t}\, x_{\ell,t}) + (1-\lambda)\,\mathbb{E}[u^*_{t+1}(x_{i,t}\, x_{j,t}\, x_{k,t}\, x_{\ell,t})]$$

$$x_{j,t} \in \{0,1\}^{\mathcal{I}_t}, (j,t) \in \mathcal{I}_t$$

$$x_{j,(t+1)} \in \{0,1\}^{\mathcal{I}_{t+1}}, (j,t+1) \in \mathcal{I}_{t+1}$$

The lower level objective is a consumer behavioral model:

$$u^*_{t+1} = \max_{C_{t+1}} \sum_{i \in A_{t+1}} \sum_{j \in R_{t+1}} \sum_{k \in E_{t+1}} \sum_{\ell \in N_{t+1}} \mathbb{E}\left(u_{C_{t+1}} \ \left(x_{i,(t+1)}\, x_{j,(t+1)}\, x_{k,(t+1)}\, x_{\ell,(t+1)}\right) \mid u_t\right)$$

where

$$u_t = \lambda \sum_{i \in A_t} \sum_{j \in R_t} \sum_{k \in E_t} \sum_{\ell \in N_t} u_{C_t} \ \left(x_{i,(t+1)}\, x_{j,(t+1)}\, x_{k,(t+1)}\, x_{\ell,(t+1)}\right) + (1-\lambda)\, p^*_{t+1}(x_{i,t}\, x_{j,t}\, x_{k,t}\, x_{\ell,t})$$

$$x_{j,t} \in \{0,1\}^{\mathcal{I}_t}, (j,t) \in \mathcal{I}_t$$

$$x_{j,(t+1)} \in \{0,1\}^{\mathcal{I}_{t+1}}, (j,t+1) \in \mathcal{I}_{t+1}$$

Lastly, there is a capacity (budget) constraint each period, of the form:

$$\sum_{(j,t+1) \in \mathcal{I}_{t+1}} \omega_{j,t+1} x_{j,t+1} \leq B_{t+1}$$

Then, we can combine these to form the QuKP:

$$\begin{aligned}
\max_{C_{t+1}} \quad & \sum_{i\in A_{t+1}} \sum_{j\in R_{t+1}} \sum_{k\in E_{t+1}} \sum_{\ell\in N_{t+1}} \mathbb{E}\left(p_{C_{t+1}}\, x_{i,(t+1)}\, x_{j,(t+1)}\, x_{k,(t+1)}\, x_{\ell,(t+1)} \mid p_t\right) \\
\text{s.t.} \quad & \mathbb{E}[p_{t+1}(\cdot)] \geq \alpha_t \\
& \mathbb{E}[u^*_{t+1}(\cdot)] \geq \beta_t \\
& \sum_{(j,t+1)\in\mathcal{I}_{t+1}} \omega_{j,t+1} x_{j,t+1} \leq B_{t+1} \\
& x_{j,t} \in \{0,1\}^{\mathcal{I}_t}, (j,t) \in \mathcal{I}_t \\
& x_{j,(t+1)} \in \{0,1\}^{\mathcal{I}_{t+1}}, (j,t+1) \in \mathcal{I}_{t+1}
\end{aligned} \tag{QuKP}$$

From the expression of the DP stage cost

$$J_k(x_k) = \min_{u_k\in U_k(x_k)} \mathbb{E}_{w_k}\left\{g_k(x_k,u_k,w_k) + J_{k+1}(f_k(x_k,u_k,w_k))\right\}$$
$$k = 0,1,\ldots,N-1$$

we estimate future cost by optimizing lower-level objective $u^*_{t+1}(x_{i,t}\, x_{j,t}\, x_{k,t}\, x_{\ell,t})$, and can choose how much we want to rely on this prediction model, versus our observed past actions and profits.

Note, one could formulate the two-variable objective in a nonlinear knapsack problem as

$$\max \sum_{i=1}^{n} p_i x_i + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i\neq j}} P_{ij} x_i x_j$$

to make the synergy profit more explicit. There would be $\binom{4}{2} = 6$ two-variable summations, $\binom{4}{3} = 4$ three-variable summations, and one four-variable quartic summation (the one seen in the model above) that contribute to each periods profits, as well as the profits derived from the singleton items. We prefer (QuKP) for its compactness. Without loss of generality we assume that $\max_{j\in\mathcal{I}} \omega_j \leq c < \sum_{j\in\mathcal{I}} \omega_j$ and that the profit matrix is constructed to avoid double counting of rewards/profits.

Incentive and participation constraints lend themselves towards the effective "typing" of individuals and groups in the intervention scheme, which can be technically described as an *inverse optimization* procedure for the elicitation of the agents' true "type", which is their efficiency score towards learning/adaptation, mainly, and controls the rate at which the agent (or group of agents) learn and apply their learning in the intervention environment. There are also other "types" that are of interest, and are domain specific, such as the intensity of the agent's impact on nodes close to it in the social network, which can be used to add precision to network-based intervention actions (on the part of the principal).

Inherently, consumer behavioral models are utility maximizing, and we will stick to this regime, assuming the principal and agent, according to the PAM, negotiate the expected effort level of the agent and the nature of compensation for the *output produced by that effort*, which in the context of Global Poverty Alleviation & International Development, should include capacity-building efforts (i.e., efforts to expand the capabilities of the agents' production systems, which do not in themselves produce output, things like hirings, updated supplier lists and relationships, obtainment of more production space or resources, etc.).

Motivating example: Food Interventions

Here we reduce our model to the main model presented in this paper by Levi et al. [60]. Their model is a bilevel optimization method whose main equation is written as:

$$\begin{aligned} \max_{\beta,\delta,\nu} h^*/T(h^*, u^*; v(\nu), d(\delta)) & \\ \text{s.t.} \quad \beta + \delta + \nu \leq U & \\ \beta, \delta, \nu \geq 0 & \\ h^*, u^* = \arg\max M^c(B(\beta), d(\delta), v(\nu)) & \end{aligned} \tag{M^g}$$

where U is the governments budget per customer. The functions $v(\cdot)$, $d(\cdot)$, and $B(\cdot)$ link the government's monetary investments in each intervention to (the agents') value of nutrition, disutility (time cost), and food budget, respectively. Here the customer (agent's) utility function, our *consumer behavioral model*, $M^c(B, d, v)$ is given in §2.1 of the Levi paper. The principal's objective of maximizing healthy foods, given by the objective function in their model, can be represented by requiring the quality of some resource category, like resources spent on healthy eating, to be true, that is, valued "1".

Chapter 6

Conclusions and Future Work

Here we conclude this dissertation work with a discussion on shared costs and savings models, which are the main management tools to track the dynamic progression of our intervention framework. Policy and practice recommendations, generated by the solutions to the simulations we conduct using our game theoretic models, will be built on this. Finally, we conclude with a note on pedagogical contributions to the field of development engineering and engineering ethics, and leave a future note towards the importance of simulation and computing technology, training, and strategy going forward.

6.1 Cost Savings Models

At the national, state, and city level in the United States, large amounts of cost savings and shared benefits will be gained if we can discover and implement the best *characteristic states* for individuals and communities on an array of issues that cause and reify poverty and other social ills. In short, the impact one can have with intervention models on a small scale may be negligible given the investment necessary to intervene and optimally track the intervention pathway; however, if we can understand common themes across hundreds of cities, dozens of states, and up to the entirety of sovereign nations, we can both permanently eradicate socially detrimental practices and behaviors, and produce large national and subnational cost savings and shared benefits across peoples, that will be inherited by future generations of Americans and international citizens.

Definitions here are from the business world. *Cost reduction* is a raw dollar-savings approach with the primary objective of cutting necessary expenses by changing the scope of maintenance services without negatively impacting the end result. For example: a retail store might reduce its monthly floor cleaning service to quarterly during the down season. That methodology, however, does not account for the non-hourly expenditures associated with maintenance, such as additional service trip costs, emergency repair fees, and overtime charges. That means potentially bloated true costs when all is said and done. In contrast, *cost savings* is a holistic strategy to maintenance budgeting that looks at more than just the upfront dollar amount of services. Rather than cutting those monthly floor cleaning services, for example, a retailer would consider alternative measures to compress service costs—such as leveraging a service provider network instead of standalone contractors—while still fulfilling

their objectives. This is also known as a total cost approach.

Similar to the cost savings models is the idea of *shared savings* programs, which will be integral to the success of OIT. For example, the Medicare Shared Savings Program (MSSP) offers providers and suppliers (e.g., physicians, hospitals, and others involved in patient care) an opportunity to create an Accountable Care Organization (ACO). An ACO agrees to be held accountable for the quality, cost, and experience of care of an assigned Medicare fee-for-service (FFS) beneficiary population. The Shared Savings Program has different tracks that allow ACOs to select an arrangement that makes the most sense for their organization. The Shared Savings Program is an important innovation for moving the Centers for Medicare & Medicaid Services' (CMS') payment system away *from volume and toward value and outcomes*. Towards this end, the referenced "artistic integration of independent systems" with OIT hinges on the creation and maintenance of optimal subsystems which work together to consistently measure and promote shared savings character states, and thus future interventions.

6.2 Future Work

In addition to our mathematical and statistical training, there is a need for OR practitioners to return to the early days of the field, where they were trained in computation techniques and were knowledgeable of computer structures. With the advance of scientific and high performance computing, and with the scale of the models we hope to simulate, appropriate parallelization, memory storage, usage, and transfer, and cloud/server computing strategies will be important for the simulation side of OIT. We hoped to have a subsection with a fuller treatment on simulation and computing concerns for OIT and similar AI/ML for social good practitioners, but will have to leave this to future work.

6.3 Personal Reflections on Engineering Ethics and other Pedagogical Contributions

Following a course in engineering ethics, the thing that stuck closest to my mind and heart was the concept of *utilitarianism*. The concept of deriving the most good for the largest amount of people with our technical engineering solutions is another way, perhaps a better way, to express the way I feel about the future of engineering science. I consider myself an engineering interventionist. I use statistical learning theory as the basis to understand how we train systems, and will prescribe methods and policies that help us train entire human systems towards a well-defined objective. This new age, this new researcher, that uses her skill in all disciplines to intervene in society's most marginalized communities, where the government, by way of the tragedy of the commons, fails them, is my passion in engineering.

This semester assured me that engineering ethics is the right field inside of academia to address philosophical questions on where and what we should study as engineers. Community-based participatory research (CBPR) seems to be the right term outside of the academia. Putting these two together, I feel very strongly about the way engineers should choose what they work on (ensuring that those who need the most help get the most constructive help),

how they should consult stakeholders (human-centered design), and how they should use their position in society, to bring about a higher quality of physical and social life amongst the masses.

I recall a conversation I had with a few colleagues when we were in Uganda over winter break, concerning this concept of the outsider ascribing to some society their "quality of life". It is true that we should always be culturally humble [101]; however, the prospects of physical quality of life, in my opinion, must be reached by World consensus. To say a people are happy in their society where 5% of the country is electrified, as in Uganda, is a disservice from an engineering ethics lens. How would they know about desalination, steam purification, hydroelectricity, etc., things that they could readily use and train their own engineers to use, if we do not take a leading role in society as engineers? Towards this end, and back to the discourse, utilitarianism is a great way to think about my passion, and to explain it to others. Are engineers doing the best they can for the greatest amount of people in the World? Or do our engineering innovations in the West only serve to further worldwide disdain and disillusionment with Western technology and power.

Bibliography

[1] Mónica Hernández Alava. "Growth dynamics: an empirical investigation of output growth using international data". PhD thesis. Economics, 2002.

[2] James Andreoni. "Impure altruism and donations to public goods: A theory of warm-glow giving". In: *The economic journal* 100.401 (1990), pp. 464–477.

[3] Douglas W Arner, Janos Barberis, and Ross P Buckley. "The evolution of Fintech: A new post-crisis paradigm". In: *Geo. J. Int'l L.* 47 (2015), p. 1271.

[4] Sanjeev Arora et al. "Proof verification and the hardness of approximation problems". In: *Journal of the ACM (JACM)* 45.3 (1998), pp. 501–555.

[5] Kenneth J Arrow. "Social choice and individual values". In: *Social Choice and Individual Values*. Yale university press, 2012.

[6] Anil Aswani. "Low-rank approximation and completion of positive tensors". In: *SIAM Journal on Matrix Analysis and Applications* 37.3 (2016), pp. 1337–1364.

[7] Claudine Badue et al. "Self-driving cars: A survey". In: *Expert Systems with Applications* 165 (2021), p. 113816.

[8] Abhijit Vinayak Banerjee, Esther Duflo, and Michael Kremer. "The influence of randomized controlled trials on development economics research and on development policy". In: *The state of Economics, the state of the world* (2016), pp. 482–488.

[9] Boaz Barak and Ankur Moitra. "Noisy tensor completion via the sum-of-squares hierarchy". In: *Conference on Learning Theory*. PMLR. 2016, pp. 417–445.

[10] P. Bartlett and S. Mendelson. "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results". In: *J. Mach. Learn. Res.* (2002).

[11] Peter L Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482.

[12] Julie Beaulac, Elizabeth Kristjansson, and Steven Cummins. "Peer reviewed: A systematic review of food deserts, 1966-2007". In: *Preventing chronic disease* 6.3 (2009).

[13] Timo Berthold, Gregor Hendel, and Thorsten Koch. "From feasibility to improvement to proof: three phases of solving mixed-integer programs". In: *Optimization Methods and Software* 33.3 (2018), pp. 499–517.

[14] Dimitri P Bertsekas et al. *Dynamic programming and optimal control*. Vol. 1. 2. Athena scientific Belmont, MA, 1995.

[15] Rajendra Bhatia. *Matrix analysis*. Springer-Verlag, 1997.

[16] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[17] Gábor Braun et al. "Blended conditonal gradients". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 735–743.

[18] Sonia Bridge and Ken Deeley. "Simulating the Ramsey-Cass-Koopmans Model Using MATLAB and Simulink". In: *Mathworks* (2016).

[19] Benjamin Bruder et al. "Regularization of portfolio allocation". In: *Available at SSRN 2767358* (2013).

[20] Caleb Bugg and Anil Aswani. "Logarithmic sample bounds for Sample Average Approximation with capacity-or budget-constraints". In: *Operations Research Letters* 49.2 (2021), pp. 231–238.

[21] Caleb Bugg, Chen Chen, and Anil Aswani. *Nonnegative Tensor Completion via Integer Optimization*. 2021. DOI: 10.48550/ARXIV.2111.04580. URL: https://arxiv.org/abs/2111.04580.

[22] Francesco Caselli and Jaume Ventura. "A representative consumer theory of distribution". In: *American Economic Review* 90.4 (2000), pp. 909–926.

[23] Venkat Chandrasekaran et al. "The convex geometry of linear inverse problems". In: *Foundations of Computational mathematics* 12.6 (2012), pp. 805–849.

[24] Charles West Churchman. *Challenge to reason*. McGraw-Hill New York, 1968.

[25] Ronald Harry Coase. *The firm, the market, and the law*. University of Chicago press, 2012.

[26] Daniel Cohen. "The HIPC initiative: true and false promises". In: *International Finance* 4.3 (2001), pp. 363–380.

[27] Justin Dauwels et al. "Handling missing data in medical questionnaires using tensor decompositions". In: *2011 8th International Conference on Information, Communications Signal Processing*. 2011, pp. 1–5. DOI: 10.1109/ICICS.2011.6174300.

[28] RF Drenick. "Multilinear programming: Duality theories". In: *Journal of optimization theory and applications* 72.3 (1992), pp. 459–486.

[29] Jitka Dupacová and Roger Wets. "Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems". In: *Annals of Statistics* 16.4 (1988), pp. 1517–1549.

[30] James T Edwards. *Organizational behavior, efficiency, and dynamics in non-profit markets: Evidence from transitional housing*. Tech. rep. Working Paper, 2013.

[31] UK. Essays. *Ramsey-Cass-Koopmens (RCK) Model — Analysis*. 2013. URL: https://www.ukessays.com/essays/economics/ramseycasskoopmens-rck-model-analysis-9339.php?vref=1 (visited on 11/22/2017).

[32] Michael Fitzgerald et al. "Embracing digital technology: A new strategic imperative". In: *MIT sloan management review* 55.2 (2014), p. 1.

[33] S. Friedland and L.-H. Lim. "Computational complexity of tensor nuclear norm". In: (2014). Submitted.

[34] Shmuel Friedland and Lek-Heng Lim. "The computational complexity of duality". In: *SIAM Journal on Optimization* 26.4 (2016), pp. 2378–2393.

[35] Silvia Gandy, Benjamin Recht, and Isao Yamada. "Tensor completion and low-n-rank tensor recovery via convex optimization". In: *Inverse problems* 27.2 (2011), p. 025010.

[36] Robert Gibbons et al. "A primer in game theory". In: (1992).

[37] Vincent Guigues, Anatoli Juditsky, and Arkadi Nemirovski. "Non-asymptotic confidence bounds for the optimal value of a stochastic program". In: *Optimization Methods and Software* 32.5 (2017), pp. 1033–1058.

[38] Marjaana Gunkel. "Incentive Design". In: *The Palgrave Encyclopedia of Strategic Management*. Ed. by Mie Augier and David J. Teece. London: Palgrave Macmillan UK, 2018, pp. 700–701. ISBN: 978-1-137-00772-8. DOI: 10.1057/978-1-137-00772-8_693. URL: https://doi.org/10.1057/978-1-137-00772-8_693.

[39] Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual*. 2021. URL: https://www.gurobi.com.

[40] Pierre Hansen. "Methods of nonlinear 0-1 programming". In: *Annals of Discrete Mathematics*. Vol. 5. Elsevier, 1979, pp. 53–70.

[41] C. Hillar and L.-H. Lim. "Most Tensor Problems Are NP-Hard". In: *J. ACM* 60.6 (Nov. 2013), 45:1–45:39. ISSN: 0004-5411. DOI: 10.1145/2512329. URL: http://doi.acm.org/10.1145/2512329.

[42] W. Hoeffding. "Probability inequalities for sums of bounded random variables". In: *JASA* 58.301 (1963), pp. 13–30.

[43] John H Holland. "The global economy as an adaptive process". In: *The economy as an evolving complex system*. CRC Press, 2018, pp. 117–124.

[44] Anil K Jain. "Data clustering: 50 years beyond K-means". In: *Pattern recognition letters* 31.8 (2010), pp. 651–666.

[45] Michael Jünger, Gerhard Reinelt, and Giovanni Rinaldi. "The traveling salesman problem". In: *Handbooks in operations research and management science* 7 (1995), pp. 225–330.

[46] S. Kakade and A. Tewari. "Rademacher Composition and Linear Prediction". Lecture 17 notes for 'CMSC 35900 Learning Theory'. 2008.

[47] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. "Regularization techniques for learning with matrices". In: *Journal of Machine Learning Research* 13.Jun (2012), pp. 1865–1890.

[48] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. "On the complexity of linear prediction: Risk bounds, margin bounds, and regularization". In: *NeurIPS*. 2009, pp. 793–800.

[49] Tapas Kanungo et al. "An efficient k-means clustering algorithm: Analysis and implementation". In: *IEEE transactions on pattern analysis and machine intelligence* 24.7 (2002), pp. 881–892.

[50] Sujin Kim, Raghu Pasupathy, and Shane G Henderson. "A guide to sample average approximation". In: *Handbook of Simulation Optimization*. Springer, 2015, pp. 207–243.

[51] Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de-Mello. "The sample average approximation method for stochastic discrete optimization". In: *SIAM Journal on Optimization* 12.2 (2002), pp. 479–502.

[52] T. Kolda and B. Bader. "Tensor Decompositions and Applications". In: *SIAM Review* 51.3 (2009), pp. 455–500.

[53] Aryeh Kontorovich. "Concentration in unbounded metric spaces and algorithmic stability". In: *ICML*. 2014, pp. 28–36.

[54] Tomi Kortela. "Growth and precautionary saving in the Ramsey-Cass-Koopmans economy". In: ().

[55] Jean-Jacques Laffont and David Martimort. *The theory of incentives*. Princeton university press, 2009.

[56] Darius Lakdawalla and Tomas Philipson. "The nonprofit sector and industry performance". In: *Journal of Public Economics* 90.8-9 (2006), pp. 1681–1698.

[57] Guillaume Lecué et al. "Empirical risk minimization is optimal for the convex aggregation problem". In: *Bernoulli* 19.5B (2013), pp. 2153–2166.

[58] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991. ISBN: 9783540520139.

[59] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer-Verlag, 1991.

[60] Retsef Levi, Elisabeth Paulson, and Georgia Perakis. "Optimal interventions for increasing healthy food consumption among low income households". In: (2019).

[61] Xin Li, Bahadir Gunturk, and Lei Zhang. "Image demosaicing: A systematic survey". In: *Visual Communications and Image Processing 2008*. Vol. 6822. International Society for Optics and Photonics. 2008, 68221J.

[62] Hongcheng Liu, Charles Hernandez, and Hung Yi Lee. "Regularized Sample Average Approximation for High-Dimensional Stochastic Optimization Under Low-Rankness". In: *arXiv preprint arXiv:1904.03453* (2019).

[63] Hongcheng Liu et al. "Sample average approximation with sparsity-inducing penalty for high-dimensional stochastic programming". In: *Mathematical Programming* (2018), pp. 1–40.

[64] Ji Liu et al. "Tensor completion for estimating missing values in visual data". In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), pp. 208–220.

[65] James Luedtke and Shabbir Ahmed. "A sample approximation approach for optimization with probabilistic constraints". In: *SIAM Journal on Optimization* 19.2 (2008), pp. 674–699.

[66] Harry Markowitz. "Portfolio selection". In: *Journal of Finance* 7.1 (1952), pp. 77–91.

[67] Pascal Massart. "Some applications of concentration inequalities to statistics". In: *Annales de la faculté des sciences de Toulouse Sér. 6* 9.2 (2000), pp. 245–303.

[68] Colin McDiarmid. "On the method of bounded differences". In: *Surveys in Combinatorics* 141.1 (1989), pp. 148–188.

[69] Yonatan Mintz and Anil Aswani. "Polynomial-time approximation for nonconvex optimization problems with an L1-constraint". In: *IEEE CDC*. IEEE. 2017, pp. 682–687.

[70] Andrea Montanari and Nike Sun. "Spectral algorithms for tensor completion". In: *Communications on Pure and Applied Mathematics* 71.11 (2018), pp. 2381–2425.

[71] Cun Mu et al. "Square deal: Lower bounds and improved relaxations for tensor recovery". In: *International conference on machine learning*. PMLR. 2014, pp. 73–81.

[72] Roy E Murphy. *Adaptive Processes in Economic Systems by Roy E Murphy*. Elsevier, 2000.

[73] Carmeliza Navasca, Lieven De Lathauwer, and Stefan Kindermann. "Swamp reducing technique for tensor decomposition". In: *2008 16th European Signal Processing Conference*. IEEE. 2008, pp. 1–5.

[74] Arkadi Nemirovski. "Topics in non-parametric statistics". In: *Ecole d'Eté de Probabilités de Saint-Flour* 28 (2000), p. 85.

[75] Roberto I Oliveira and Philip Thompson. "Sample average approximation with heavier tails I: non-asymptotic bounds with weak assumptions and stochastic constraints". In: *arXiv:1705.00822* (2017).

[76] Roberto I Oliveira and Philip Thompson. "Sample average approximation with heavier tails ii: localization in stochastic convex optimization and persistence results for the lasso". In: *arXiv:1711.04734* (2017).

[77] Manfred Padberg. "The boolean quadric polytope: some characteristics, facets and relatives". In: *Mathematical programming* 45.1-3 (1989), pp. 139–172.

[78] Christos H Papadimitriou and Mihalis Yannakakis. "Optimization, approximation, and complexity classes". In: *Journal of computer and system sciences* 43.3 (1991), pp. 425–440.

[79] Mark Pauly and Michael Redisch. "The not-for-profit hospital as a physicians' cooperative". In: *The American Economic Review* 63.1 (1973), pp. 87–99.

[80] David Pisinger. "The quadratic knapsack problem—a survey". In: *Discrete applied mathematics* 155.5 (2007), pp. 623–648.

[81] David Pisinger, Anders Bo Rasmussen, and Rune Sandvik. "Solution of Large-sized Quadritic Knapsack Problems Through Aggressive Reduction". In: *Diku Tech Report 04/11*. University of Copenhagen, Institute of Computer Science, 2004.

[82] Y. Qi, P. Comon, and L.-H. Lim. "Uniqueness of Nonnegative Tensor Approximations". In: *arXiv preprint arXiv:1410.8129* (2014).

[83] Yang Qi, Pierre Comon, and Lek-Heng Lim. "Uniqueness of nonnegative tensor approximations". In: *IEEE Transactions on Information Theory* 62.4 (2016), pp. 2170–2183.

[84] Nikhil Rao, Parikshit Shah, and Stephen Wright. "Forward–Backward Greedy Algorithms for Atomic Norm Regularization". In: *IEEE Transactions on Signal Processing* 63.21 (2015), pp. 5798–5811. DOI: 10.1109/TSP.2015.2461515.

[85] Holger Rauhut and Željka Stojanac. "Tensor theta norms and low rank recovery". In: *Numerical Algorithms* 88.1 (2021), pp. 25–66.

[86] R. Rockafellar and R. Wets. *Variational Analysis*. Springer, 2009.

[87] Susan Rose-Ackerman. "Altruism, nonprofits, and economic theory". In: *Journal of economic literature* 34.2 (1996), pp. 701–728.

[88] Johannes O Royset. "On sample size control in sample average approximations for solving smooth stochastic programs". In: *Computational Optimization and Applications* 55.2 (2013), pp. 265–309.

[89] Johannes O Royset and Roberto Szechtman. "Optimal budget allocation for sample average approximation". In: *Operations Research* 61.3 (2013), pp. 762–776.

[90] A.P. Ruszczyński and A. Shapiro. *Stochastic Programming*. Handbooks in operations research and management science. Elsevier, 2003.

[91] UN SDG. "Sustainable development goals". In: *The energy progress report. Tracking SDG* 7 (2019).

[92] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[93] Alexander Shapiro. "Monte Carlo sampling methods". In: *Handbooks in operations research and management Science* 10 (2003), pp. 353–425.

[94] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[95] Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. "A review on bilevel optimization: from classical to evolutionary approaches and applications". In: *IEEE Transactions on Evolutionary Computation* 22.2 (2017), pp. 276–295.

[96] Stephen A Smith. *Contract theory*. OUP Oxford, 2004.

[97] Qingquan Song et al. "Tensor completion algorithms in big data analytics". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13.1 (2019), pp. 1–48.

[98] Nathan Srebro and Adi Shraibman. "Rank, trace-norm and max-norm". In: *International Conference on Computational Learning Theory*. Springer. 2005, pp. 545–560.

[99] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. "Smoothness, Low Noise and Fast Rates". In: *Advances in Neural Information Processing Systems*. 2010, pp. 2199–2207.

[100] V. Sudakov. "Gaussian random processes and solid angle measures in Hilbert space". In: *Doklady Akademii Nauk SSSR* 197.1 (1971), pp. 43–45.

[101] Melanie Tervalon and Jann Murray-Garcia. "Cultural humility versus cultural competence: A critical distinction in defining physician training outcomes in multicultural education". In: *Journal of health care for the poor and underserved* 9.2 (1998), pp. 117–125.

[102] Alexandre B Tsybakov. "Optimal rates of aggregation". In: *Learning theory and kernel machines*. Springer, 2003, pp. 303–313.

[103] Bram Verweij et al. "The sample average approximation method applied to stochastic routing problems: a computational study". In: *Computational Optimization and Applications* 24.2-3 (2003), pp. 289–333.

[104] Burton Allen Weisbrod. *The nonprofit economy*. Harvard University Press, 2009.

[105] Andrzej P Wierzbicki. "A mathematical basis for satisficing decision making". In: *Mathematical modelling* 3.5 (1982), pp. 391–405.

[106] Wikipedia contributors. *Control theory — Wikipedia, The Free Encyclopedia*. [Online; accessed 22-April-2022]. 2022. URL: https://en.wikipedia.org/w/index.php?title=Control_theory&oldid=1083443871.

[107] Wikipedia contributors. *Externality — Wikipedia, The Free Encyclopedia*. [Online; accessed 19-April-2022]. 2022. URL: https://en.wikipedia.org/w/index.php?title=Externality&oldid=1080380670.

[108] Wikipedia contributors. *Tragedy of the commons — Wikipedia, The Free Encyclopedia*. [Online; accessed 19-April-2022]. 2022. URL: https://en.wikipedia.org/w/index.php?title=Tragedy_of_the_commons&oldid=1083415832.

[109] HS Witsenhausen. "A simple bilinear optimization problem". In: *Systems & control letters* 8.1 (1986), pp. 1–4.

[110] Ming Yuan and Cun-Hui Zhang. "Incoherent tensor norms and their applications in higher order tensor completion". In: *IEEE Transactions on Information Theory* 63.10 (2017), pp. 6753–6766.

[111] Ming Yuan and Cun-Hui Zhang. "On tensor completion via nuclear norm minimization". In: *Foundations of Computational Mathematics* 16.4 (2016), pp. 1031–1068.

[112] Xiaoqin Zhang et al. "Robust low-rank tensor recovery with rectification and alignment". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2019), pp. 238–255.

Appendix A

Glossary

Here we provide definitions as I consider them, for engineering, economic, and academic concepts integral to Optimal Intervention Theory. Some definitions are taken from dictionaries and other academic works, while some are personal to the author. Each definition here will be used in-context in the text of the research manifest. A thorough reading of this word guide before, or apart from, the rest of the document should provide the reader with clarity towards the work, and therefore is an independent abstract in itself. Of import here are the following definitions: Iterative engineering, Optimal Intervention Theory, (the) Social Economy, and the Principal Agent Model.

A.1 List of terms

- Bilevel program
- Consumer Behavioral Model
- Contract Theory
- Control Theory
- Digital Technology
- Dynamic Programming
- Economic Decision Analysis (EDA)
- Game Theory
- Healthy and Active Lifestyles (HAL)
- Incentive Design
- Iterative Engineering
- Machine Learning (ML)
- Mechanism Design

- Optimal Intervention Theory (OIT)
- Positive Externality
- Principal Agent Model (PAM)
- Reinforcement Learning
- Shared Cost/Savings
- Social Economy
- Social Welfare Functions (SWF's)
- Socially Beneficial Goods and Services (SBGs)
- Statistic
- Tractable Solutions
- Tragedy of the Commons
- Transaction Costs
- United Nations' Sustainable Development Goals (UN SDGs)

A.2 Definitions

Definition 1 (Bilevel Program)**.** *A bilevel program is an optimization problem which has an optimization problem as a constraint as well; that is, both valuating the objective function and determining feasible solutions to the constraint sets require solving an optimization problem [95]. An example taken from this reference characterizes the objective function as the "upper-level" objective function and the (optimization-based) constraint as the "lower-level" objective function. These objectives are typically "unordered", in the sense that either objective function can serve as the starting point for the optimization procedure, which alternates between to the levels of the program, using thee output of one optimization procedure as input into the other level's program.*

Definition 2 (Consumer Behavioral Model)**.** *A consumer behavior model is a theoretical framework for explaining why and how customers make purchasing decisions. The goal of consumer behavior models is to outline a predictable map of customer decisions up until conversion/adoption of the marketed good or service.*

Definition 3 (Contract Theory)**.** *Refer to [96], contract theory is the study of how people and organizations construct and develop (formal and informal) legal agreements. It analyzes how parties with conflicting interests build formal and informal contracts, even tenancy.*

Definition 4 (Control Theory)**.** *Control theory deals with the control of dynamical systems in engineered processes and machines. The objective is to develop a model or algorithm governing the application of system inputs to drive the system to a desired state, while minimizing any delay, overshoot, or steady-state error and ensuring a level of control stability; often with the aim to achieve a degree of optimality [106].*

Definition 5 (Digital Technology)**.** *Embracing digital technologies is referred to as a new strategic imperative in this work [32]. Digital technologies are electronic tools, systems, devices and resources that generate, store or process data. Well known examples include social media, online games, multimedia and mobile phones. Digital learning is any type of learning that uses technology. An important theme in the field of development engineering, and the associated economics of global poverty reduction, is the "digital transformation" of societies and peoples due to the ubiquitous presence of digital technologies.*

Definition 6 (Dynamic Programming)**.** *Using Bertsekas' introduction, DP "deals with situations where decisions are made in stages. The outcome of each decision may not be fully predictable but can be anticipated to some extent before the next decision is made. The objective is to minimize a certain cost (or maximize a certain utility)— a mathematical expression of what. is considered an undesirable outcome). Notions such as controllability, stability, observability, and convergence are paramount in the study of dynamical systems.*

Definition 7 (Economic Decision Analysis (EDA))**.** *From Georgia Tech ISyE's site, a leading systems engineering program, we have the definition "In the area of Economic Decision Analysis, faculty and student researchers at ISyE apply analytical, quantitative, and empirical techniques to planning, contracts, and decisions that involve economic, financial, or social valuations."*

Definition 8 (Game Theory)**.** *From [36], game theory is the study of multiperson decision problems, which is also described as a setting wherein "multiple strategic agents interact in a strategic environment in order to maximize their individual or group utilities". At the micro level, models of trading processes (such as bargaining and auction models) involve game theory. At an intermediate level of aggregation, labor and financial economics include game-theoretic models of the behavior of a firm in its input markets (whereas at the oligopoly level, this is done in the output markets)... Finally, at a high level of aggregation, international economics include models in which countries collude or compete in choosing tariffs or trade policies.*

Definition 9 (Healthy and Active Lifestyles (HAL))**.** *Unique to the author, this describes a thrust to provide sufficient nutritional and caloric quantity and quality, opportunities for physical exercise and rest, and sleep/rest patterns, as well as a host of other mental, emotional, and psychological health factors, so that the global citizenry can actively learn and build solutions to its societies most pressing problems. Many social interventions view adequacy of resource provision of the end goal, and this will promote health only. In order to promote active lifestyles, which are needed for citizens to prescribe and carry-out long-term interventions for themselves and their communities, we must expect and provide sufficient environments for the growth and comfort of communities of people.*

Definition 10 (Incentive Design)**.** *Defined in [38], Incentive design is a careful process of crafting a system that connects performance measurement with performance rewards, with the goal of motivating employees/citizens to perform according to the expectations of the organization/intervention and principal. In OIT, Incentive design is a means of aligning the interests of a principal and its agents.*

Definition 11 (Iterative engineering)**.** *Construction and engineering methods and practices which are built with a specific concern for iterative design; that is, they can be easily changed to accommodate improvements in system strategy or available technologies.*

An example in prototyping transportation networks includes the use of gravel lots/roads over cement/asphalt. Whereas asphalt-based roads require heavy machinery that may be expensive to procure and transport around the world, gravel lots can be used and changed easily (to greenery, reverted to dirt or clay, etc.), and in this context, represent a better iterative construction process that asphalt.

Definition 12 (Machine Learning)**.** *Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.*

Definition 13 (Mechanism Design)**.** *Mechanism design is a branch of microeconomics that explores how businesses and institutions can achieve desirable social or economic outcomes given the constraints of individuals' self-interest and incomplete information.*

Definition 14 (Optimal Intervention Theory)**.** *Optimal Intervention Theory (OIT) is a method for improving human systems based on Statistical Learning Theory (SLT), which is the basis for learning in machines.*

OIT considers the classic trade-off between statistically rigorous methods and large-scale methods, often referred to, respectively, as precision/specificity and the "curse of dimensionality". The Dynamic Programming algorithm, which solves problems with well-defined end-goals in stages, is the basis of the SLT applicable to OIT.

By improving independent systems with OIT, and integrating systems through the use of artistic engineering methods, we can solve large-scale societal issues, especially those stemming from historical and contemporary social, political, and economic misalignment due to oppressive tribalism (ie. nationalism, racism, sexism, classism, ageism, ableism, etc.) by constructing collective intelligence gathered from solving these large-scale problems at the subsystem scale (i.e. Community-Based Operations Research), such that we intervene optimally in new communities based on gathered best practices.

The ability to solve problems (i.e. intervene) at the community-scale can be thought of as exploration, and the use of best practices in subsequent interventions can be looked at as exploitation, from the classic RL (Reinforcement Learning) paradigm.

Definition 15 (Reinforcement Learning)**.** *Reinforcement learning is a machine learning training method based on rewarding desired behaviors and/or punishing undesired ones. In general, a reinforcement learning agent is able to perceive and interpret its environment, take actions and learn through trial and error.*

Definition 16 (Positive Externality)**.** *In economics, an externality or external cost is an indirect cost or benefit to an uninvolved third party that arises as an effect of another party's (or parties') activity. Externalities can be considered as unpriced goods involved in either consumer or producer market transactions. Air pollution from motor vehicles is one example. The cost of air pollution to society is not paid by either the producers or users of motorized transport to the rest of society [107].*

Closely related to the idea of the tragedy of the commons, and the adoption of socially beneficial goods and services (SBGs), is the idea of *positive externalities*. Positive externalities are the benefits from the adoption of SBGs, such as reduced healthcare costs for low-income communities and the nation's health infrastructure due to improved diet, exercise, and sleep/rest regimens, which reduce diet-based health problems. These positive externalities can be captured in cost savings models (see Ch. 6) and *should be included in our predictions of intervention impact*. In fact, this is the main benefit to our objective functions that we introduce, and follows the lines of work done by others in *social welfare functions*.

Definition 17 (The Principal Agent Model (PAM))**.** *The PAM describes the problem where "there exists a conflict in priorities between a person or group and the representative (i.e principal) authorized to act on their behalf. The principal-agent problem is as varied as the possible roles of principal and agent." [cite].*

This is one of the most important models and analysis tools in this work. Importantly, most expositions and solution methods to the problem solve an incentive problem for the principal towards the agent, assuming the agent to be a rational utility maximizer. We will discuss a principal agent model where the principal's sole objective is to *move the agents along an optimal pathway towards the adoption of some socially beneficial good or service*. As another note of the importance of this concept, beyond the time and space dedicated to describe different arrangements of the social economy as principal, and communities of people (esp. in the developing world) as agents, the original title of this work was "Long-term dynamic contracting under the Principal-Agent Model for the Adoption of Socially Beneficial Goods.

Definition 18 (Shared Cost/Savings Models)**.** *These are specific models in economics, and games in Game Theory, which describe the situation where the costs of a public good or service is reduced whenever more users adopt the good/service. This is important to Optimal Intervention Theory (OIT), as many private-public funding schemes must be funded at-scale; that is, there must be predictive evidence that the good or service marketed has an increasingly impactful effect on the targeted population, while simultaneously decreasing the cost for the community of people to access, learn about, and resource the common good or service, in order to make investment in the projects worthwhile (private capital investment requires a (long-run) rate of return).*

Definition 19 (The Social Economy)**.** *In the U.S. context, the social economy is comprised of:*

1. *The Academy (Institutions for Higher Education)*

2. *Private Philanthropy*
3. *The Government, at all levels*
4. *Community-based nonprofits*
5. *Context-agnostic emulsifiers, who have expertise in financial, engineering (and other technical), and/or evaluation techniques.*

Many pathways/networks of interaction exist. One of import to us as academics is government funding of universities to produce research (that solves society's problems).

Definition 20 (Social Welfare Functions (SWF's))**.** *A social welfare functions (SWF), most common in welfare economics, is defined as a process or rule, which for each set of individual orderings for alternative social states (one ordering per individual), yields a corresponding ordering of alternative social states. An SWF ranks social states as less desirable, more desirable, or indifferent for every possible pair of social states.Each state is a possible complete description of the society. Inputs of the SWF include any variables considered to affect the economic welfare of the chosen society [5].*

Definition 21 (Socially Beneficial Goods and Services)**.** *An important namesake for our work, socially beneficial goods and services (SBG's) are goods and services that have both a direct benefit after adoption for the user, and a positive impact on social and socioeconomic structures. The positive impacts beyond direct benefits to the user which adopts the SBGs are called positive externalities, or social benefits. SBGs encompass development themes such as food, water, and energy security, literacy and learning development, financial literacy/understanding and savings-based (rather than debt-based) funding of daily life, and many other issues circumscribed by the UN's Sustainable Development Goals (SDG's)."*

Definition 22 (Statistic)**.** *A statistic (singular) or sample statistic is any quantity computed from values in a sample which is considered for a statistical purpose. Statistical purposes include estimating a population parameter, describing a sample, or evaluating a hypothesis. The average (or mean) of sample values is a statistic. The term statistic is used both for the function and for the value of the function on a given sample. When a statistic is being used for a specific purpose, it may be referred to by a name indicating its purpose.*

Definition 23 (Tractable Solution)**.** *In the intervention concept, this means "interventions that last long beyond the intervention period, as measured by those indicators used to measure utility in the original intervention".*

In the mathematical and statistical world, tractable refers, loosely, to "useful/usable" models for a specific problem, that produce a usable result (as compared to intractable, where we cannot make much use. Classic economic models are often called intractable for economic practice, as opposed to theory. This is addressed by current authors as well as Murphy, in the motivation for his work.

Definition 24 (Tragedy of the Commons)**.** *An important economic notion for the adoption of SBG's is the tragedy of the commons, "in which individual users, who have open access to a resource unhampered by shared social structures or formal rules that govern access and use, act independently according to their own self-interest and, contrary to the common good of all users, cause depletion of the resource through their uncoordinated action [108]."*

This describes the main line of argument against social investments. It is a good concept and model to implore governments and socially-focused private investors to act in a way that alleviates a tragedy of the commons. This is why we have public education departments, elderly care programs (e.g. medicare, social security), and other social safety nets. Public goods such as parks, roads, and sanitation services must be maintained by a central body elected by citizens to reduce the effects of a tragedy of the commons, These central bodies are most commonly called "governments", at all levels.

Definition 25 (Transaction Costs)**.** *The well-known economist R.H. Coase in his essay "The Firm, the Market, and the Law" provides a new economic viewpoint for analysis which the author follows, which states that transaction costs should be the focus of economic theory and modelling. He defines transaction costs originally as the costs of market transactions, and writes, "In order to carry out a market transaction, it is necessary to discover who it is that one wishes to deal with, to inform people that one wishes to deal and on what terms, to conduct negotiations leading up to a bargain, to draw up the contract, to undertake the inspection needed to make sure that the terms of the contract are being observed, and so on. Dahlman crystallized the concept of transaction costs by describing them as "search and information costs, bargaining and decision costs, policing and enforcement costs"[25].*

Many, if not all, of these concepts are integral to the (dynamic) contracting arrangement of a continuous-time Principal-Agent Model, and will be described therein (in fact, the original title of this work was "Long-term dynamic contracting under the Principal-Agent Model for the Adoption of Socially Beneficial Goods"). From Ch. 1, section 7 "The way ahead", Coase writes, "I have suggested that economists need to adopt a new approach when considering economic policy... To do this it is not necessary to abandon standard economic theory, but it does mean incorporating transaction costs into the analysis, since so much that happens in the economic system is designed either to reduce transaction costs or to make possible what their existence prevents."

In the context of international development and global poverty alleviation, there are both physical (infrastructure, transportation networks, technology and energy availability, etc.) and human (political and economic disorganization, war and other sources of gainless loss, bribery and (risk-heavy) informal economies) system arrangements which cause large transaction costs for practitioners in the field. Hence, much of our theory makes the assumption that legal and political experts will, simultaneous to the development of our models and theory, lay out a theory for global peace practices and assured deterrents of war and chaos based political movements and regimes, so that investment in development projects are worthwhile (we can think of war, especially, as a stochastic (randomly occurring)"big M" cost; if it cannot be contained, no system can ever be "stable" or "controllable" (see control theory).

Definition 26 (United Nations' Sustainable Development Goals)**.** *Found on the UN's website, a simplified list of the 17 goals are given here. Since the author does not prescribe any goals for societal improvement apart from these (and focuses mainly on food and housing security for application areas), we strongly suggest the reader visits the website and reads the additional information provided, especially for the areas of your concern.*

1. *No Poverty*
2. *Zero Hunger*
3. *Good Health and Well-Being*
4. *Quality Education*
5. *Gender Equality*
6. *Clean Water and Sanitation*
7. *Affordable and Clean Energy*
8. *Decent Work and Economic Growth*
9. *Industry, Innovation, and Infrastructure*
10. *Reduced Inequalities*
11. *Sustainable Cities and Communities*
12. *Responsible Consumption and Production*
13. *Climate Action*
14. *Life Below Water*
15. *Life on Land*
16. *Peace, Justice, and Strong Institutions*
17. *Partnerships for the Goals*