The Modeling Spectrum of Data-Driven Decision Making

by

Xianglin Meng

B.A., M. of Mathematics, University of Oxford (2016)
S.M., Massachusetts Institute of Technology (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author .....
Department of Electrical Engineering and Computer Science
August 26, 2022

Certified by .....
Munther A. Dahleh
William A. Coolidge Professor
Department of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

The Modeling Spectrum of Data-Driven Decision Making

by

Xianglin Meng

Submitted to the Department of Electrical Engineering and Computer Science
on August 26, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Data-driven decision-making has become an essential part of modern life by virtue of the rapid growth in data, the massive improvements in computing power, and great progress in academic research. The range of techniques used fall broadly on the spectrum that varies from model-based to applied, depending on the problem complexity and data availability.

This thesis studies three settings that span the modeling spectrum in the contexts of digital agriculture, cell reprogramming, and pandemic policymaking. First, we investigate the problem of learning good farming practices in the framework of multi-armed bandits with expert advice. We extend the setting from finitely many experts to any countably infinite set and provide algorithms that are provably optimal. Second, we explore optimizing perturbations for cell reprogramming in batched experiments. Building upon multi-armed bandit algorithms, we propose an active learning approach that integrates deep learning and biology-based analysis. We numerically demonstrate the success of our method on gene expression data. Finally, we model the impacts of nonpharmaceutical interventions during the coronavirus disease 2019 (COVID-19) pandemic. We develop an agent-based model in order to overcome the limitations of observational data. We show that the trade-off between COVID-19 deaths and deaths of despair, dependent on the lockdown level, only exists in the socioeconomically disadvantaged population. Our model establishes effective measures for reducing disparities during the pandemic.

Thesis Supervisor: Munther A. Dahleh
Title: William A. Coolidge Professor
Department of Electrical Engineering and Computer Science

Acknowledgments

First and foremost, I would like to thank my advisor Munther Dahleh who is fundamental to my growth as a researcher in the past six years at MIT. His vision of tackling problems that are core to technical advancement and social progress has a great influence on my research taste. Munzer truly cares about his students, in the lab and beyond. I am extremely grateful to Munzer for challenging, guiding, and supporting me.

I would also like to thank my doctoral thesis committee Caroline Uhler and Devavrat Shah. I was very fortunate to be part of the first cohort of the Eric and Wendy Schmidt Center co-directed by Caroline. It was a thrilling adventure for me which initiated a fruitful collaboration with her that became an important chapter of this thesis. I have learned so much from Caroline. I have known Devavrat since my first year at MIT whose research I have always admired. Over the years, our interactions spanned almost everywhere from teaching to group meetings, journal clubs, and seminars. Devavrat's feedback on some of my work has strengthened this thesis.

I am also grateful to John Tsitsiklis who co-advised me for my Master's thesis. John's intellectual acuity and breadth of knowledge are great inspirations to me. I still vividly recall the time that we spent together bouncing ideas and digging into the details.

I would like to thank my collaborators: Tuhin Sarkar, Dalton Jones, Roberto Rigobon, Jiaqi Zhang, Louis Cammarata, James Briggs, and Feng Zhang. This thesis would not have been possible without them.

I feel very lucky to have been part of Munzer's group and call kind, fun, and smart people my labmates and friends. They are very generous with their ideas and feedback. I cherish many fond memories of our outings just as much.

I would like to thank the administrative staff of LIDS for making the community a homey atmosphere. My special thanks goes to Alina, Brian, Kate, Jennifer, Gracie, Rich, Rachel, and Carissa.

I would also like to thank my friends in Boston: Irene, Amanda, Maz, Micah,

Andrew, Will, Oriol, Cyrus, Charlotte, Yunpeng, Jen, Felix, Yuening, Victor, Jingkai, Shuo, Jie, Junang, Sisi, Jennifer, and Melody. They are my inspirations. They made Boston my home from home.

Finally, I would like to thank my family from the depths of my heart. My parents have always put education as their top priority. Although they would very much like to stay close to me, they unconditionally support me to pursue my dream no matter where it takes me. Since Marcus and I met during college, we have grown together as we ran through most part of our twenties. Marcus has been a constant source of love and support, especially during the final stretch of my journey at MIT. My heart is full of gratitude for him and my Swedish family. Thank you for always being there for me.

Funding to support this research was provided for by the Eric and Wendy Schmidt Center and the OCP Group.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data is everywhere. We leave digital footprints virtually everywhere we go, for everything we do. How we travel, what we like eating, where we live, the way we shop, etc. Data of the modern world open up new opportunities for companies, organizations, and governments to make informed decisions. With the help of massive increases in computing power, a variety of methods have been proposed for data-driven decision-making. Broadly speaking, the techniques range from entirely theoretical to purely heuristic, depending on the problem complexity and data availability. In this thesis, we study three settings that span the modeling spectrum in the contexts of digital agriculture, cell reprogramming, and pandemic policymaking.

Chapter 2 concerns the problem of multi-armed bandits with expert advice that we abstract from digital agriculture. Motivated by the need for learning good farming practices from a multitude of farmers and machine learning algorithms through sequential experiments, we extend the theoretical setting of the problem from finitely many experts to any countably infinite set and provide algorithms that have provably good performance.

In Chapter 3, we study active learning for efficient cell reprogramming, which has far-reaching implications for human disease modeling, regenerative medicine, and drug screening. We design an active learning algorithm that addresses the problem of combinatorial pure exploration under the constraints of batched experiments and a low signal-to-noise ratio. Although the complexity of this problem prohibits theo-

retical guarantees on the entire procedure, we propose a principled framework that is built upon multi-armed bandit algorithms, integrating deep learning and biology-based analysis. We demonstrate the success of our approach on gene expression data.

In Chapter 4, we investigate the impacts of nonpharmaceutical interventions (NPIs) during the coronavirus disease 2019 (COVID-19) pandemic which has affected everyone's life. The research started during the early days of COVID-19 in an effort to elucidate the differential causal effects of NPIs on different communities. We develop an agent-based model in order to overcome the limitations of observational data. Validated by US data, our findings contribute to policy modeling and evaluation for reducing inequality during a pandemic.

1.1 Summary of Individual Chapters

1.1.1 Nonstochastic Bandits with Infinitely Many Experts

We study the problem of nonstochastic bandits with expert advice, extending the setting from finitely many experts to any countably infinite set: A learner aims to maximize the total reward by taking actions sequentially based on bandit feedback while benchmarking against a set of experts. We propose a variant of Exp4.P that, for finitely many experts, enables inference of correct expert rankings while preserving the order of the regret upper bound. We then incorporate the variant into a meta-algorithm that works on infinitely many experts. We prove a high-probability upper bound of $\tilde{\mathcal{O}}\big(i^*K+\sqrt{KT}\big)$ on the regret, up to polylog factors, where i^* is the unknown position of the best expert, K is the number of actions, and T is the time horizon. We also provide an example of structured experts and discuss how to expedite learning in such case. Our meta-learning algorithm achieves optimal regret up to polylog factors when $i^* = \tilde{\mathcal{O}}\big(\sqrt{T/K}\big)$. If a prior distribution is assumed to exist for i^*, the probability of optimality increases with T, the rate of which can be fast. We conduct numerical experiments using synthetic data and simulated agricultural data to complement our theoretical findings.

1.1.2 Active Learning for Efficient Cell Reprogramming

Finding optimal interventions for cell reprogramming is challenging because of the high-dimensional state and action spaces, which make brute-force search impractical in general. We present a mathematical formulation of the problem to explain the intricacies of cell reprogramming due to the limitations of experiments. We propose the first active learning algorithm for efficient cell reprogramming that directly addresses the problem of combinatorial pure exploration under the constraints of batched experiments and a low signal-to-noise ratio. The framework combines multi-armed bandit algorithms, which have been proven to balance the exploration-exploitation trade-off optimally in simplified settings, and deep learning methods, which have enjoyed marvelous success in recognizing patterns in noisy data. The proposed algorithm also incorporates analysis based on biological knowledge. We demonstrate the success of our approach on gene expression data collected by state-of-the-art large-scale perturbation screening.

1.1.3 Impacts of COVID-19 Interventions

COVID-19 is exacerbating inequalities in the US. We build an agent-based model to elucidate the differential causal effects of NPIs on different communities and validate the results with US data. We simulate viral transmission and the consequent deterioration of economic conditions on socioeconomically disadvantaged and privileged populations. As found in data, our model shows that the trade-off between COVID-19 deaths and deaths of despair, dependent on the lockdown level, only exists in the socioeconomically disadvantaged population. Moreover, household overcrowding is a strong predictor of the infection rate. The model also yields new insights that fill in the gaps of our data analysis. While subsidization narrows the socioeconomic gap in deaths of despair, the combination of testing and contact tracing alone is effective at reducing disparities in both types of death. Our results contribute to policy modeling and evaluation for reducing inequality during a pandemic.

Chapter 2

Nonstochastic Bandits with Infinitely Many Experts

2.1 Introduction

Early work on the multi-armed bandit problem commonly studied settings where the rewards of each arm are stochastically generated from some unknown distribution [11, 85, 114]. In general, such statistical assumptions are difficult to validate or inapproriate for some applications such as packet transmission in communication networks [12, 13]. The problem of nonstochastic bandits, first investigated in [12, 13], makes no statistical assumptions about how the rewards are generated.

A setting of the nonstochastic bandit problem allows for incorporating expert advice. The *learner* interacts with an *adversary* over a time horizon T as follows. At each time, the adversary sets the rewards for K actions and keeps them secret. After getting every expert's advice on the probability of choosing each action, the learner combines the advice and samples an action. Finally, the learner observes only the reward of the action chosen, and the game repeats. The learner's goal is to minimize *regret*, which is the gap between the total reward gained and the expected total reward of the best expert i^* who is unknown a priori.

The framework described is a general one. First, there is no assumption about the generation of rewards except that the adversary is *oblivious*. In other words, the

adversary's choices are independent of the learner's strategy. Equivalently, all rewards can be assigned before the game starts, and the learner only observes the rewards of chosen actions sequentially. Second, we do not restrict or assume knowledge of how the experts come up with their advice. Third, experts can give deterministic advice.

The problem of bandits with expert advice is not only a natural model for numerous real-world applications, such as selecting and pricing online advertisements [95], but also important from a theoretical perspective. Contextual bandits can be framed as a bandits with expert advice problem by introducing policies that map a context to a probability distribution over actions [95, 4]. Bandits with expert advice are also closely related to online model selection where experts correspond to model classes [34, 55, 56].

Prior work on nonstochastic bandits with expert advice typically assumes the number of experts to be *finite* [12, 13, 95, 22, 101]. The **exp**onential-weight algorithm for **exp**loration and **exp**loitation using **exp**ert advice (Exp4), introduced by [12, 13], has a regret upper bound of $\mathcal{O}\big(\sqrt{KT\ln N}\big)$ *in expectation*, where N is the number of experts. This upper bound almost matches the lower bound $\Omega(\sqrt{(KT\ln N)/\ln K})$ derived by [3] for the expected regret when $\ln N \leq T\ln K$. However, Exp4 does not satisfy a similar regret guarantee *with high probability* due to the large variance of its estimates. Algorithms with high-probability guarantees are preferred for domains that need reliable methods, but such algorithms require delicate analysis [22, 101]. The Exp4.P algorithm, a variant of Exp4 proposed by [22], satisfies a regret upper bound of $\mathcal{O}\big(\sqrt{KT\ln(N/\delta)}\big)$ with probability at least $1-\delta$. This bound can be improved by a constant factor by avoiding explicit exploration [101].

We study the problem of nonstochastic bandits with infinitely many experts. Our main question is: Can the learner perform almost as well as the globally best expert i^* of a countably infinite set while only querying a finite number of experts? This question is motivated by challenges encountered in practical situations where it is unfeasible to seek advice from all experts all the time [120]. For search engine advertising, a company may need to choose among a multitude of schemes some of which also involve hyperparameter tuning [95]. As another example, there are often a myr-

iad of features that can be used for online recommendation systems. Some features tend to be more informative than others, but their relevance is normally unknown a priori. We can transform this problem into bandits with expert advice where each expert corresponds to a model class in a certain feature space. The number of experts can be extremely large due to the combinatorial nature. In contrast to the large number of experts available, it is desirable to query only some of them each time in consideration of computational constraints.

2.1.1 Problem Formulation

Let $\mathbb{Z}_+$ be the set of strictly positive integers. For $N \in \mathbb{Z}_+$, we define $[N] \triangleq \{1, 2, \ldots, N\}$. Let $T \in \mathbb{Z}_+$ be the *time horizon*. Let $\mathcal{A}$ be a set of *actions* where $|\mathcal{A}| = K < \infty$.

At each time $t \in [T]$, the adversary first sets a reward vector $r(t) \in [0,1]^K$ where $r_a(t)$ is the *reward* of action a. Each expert $i \in \mathbb{Z}_+$ then gives their *advice* $\xi^i(t)$, which is a probability vector over $\mathcal{A}$. After querying a finite subset of the experts' advice but not the rewards, the learner then samples an action $a(t)$. Finally, the learner receives the reward $r_{a(t)}(t)$ and no other information. The game proceeds to time $t+1$ and finishes after T time steps. The learner's goal is to combine the experts' advice such that the total reward is close to a benchmark, which we will define shortly.

Let $y_i(t) \triangleq \sum_{a \in \mathcal{A}} \xi_a^i(t) r_a(t)$ be the expected reward of expert i at time t. For any time interval $\mathcal{T} \subset \mathbb{Z}_+$ such that $|\mathcal{T}| < \infty$, we denote the expected total reward of expert i during $\mathcal{T}$ as $R_i(\mathcal{T}) \triangleq \sum_{t \in \mathcal{T}} y_i(t)$. We define the best expert $i^*(\mathcal{I}; \mathcal{T})$ of a subset $\mathcal{I} \subseteq \mathbb{Z}_+$ during $\mathcal{T}$ as the one with the lowest index that has the highest total reward in expectation,[1] namely, $i^*(\mathcal{I}; \mathcal{T}) \triangleq \min\{\operatorname{argmax}_{i \in \mathcal{I}} R_i(\mathcal{T})\}$. The learner's *regret* with respect to $i^*(\mathcal{I}; \mathcal{T})$ is

$$\operatorname{Regret}(\mathcal{T}; \mathcal{I}) \triangleq R_{i^*(\mathcal{I};\mathcal{T})}(\mathcal{T}) - \sum_{t \in \mathcal{T}} r_{a(t)}(t).$$

For simplicity of notation, let $\operatorname{Regret}(T) \triangleq \operatorname{Regret}([T]; \mathbb{Z}_+)$ and $i^* \triangleq i^*(\mathbb{Z}_+; [T])$. The

[1] If $\max_{i \in \mathcal{I}} R_i(\mathcal{T})$ does not exist, we define $i^*(\mathcal{I}; \mathcal{T}) = \infty$ and $R_{i^*(\mathcal{I};\mathcal{T})}(\mathcal{T}) = \sup_{i \in \mathcal{I}} R_i(\mathcal{T})$.

learner's goal is to minimize Regret(T), the regret with respect to the *globally best expert* i^* for the time horizon considered.

2.1.2 Contributions

For the general case without any assumption about the experts, we propose an algorithm called **Be**st **E**xpert **S**earch (BEES) and provide theoretical guarantees on its performance. BEES runs a subroutine called Exp4.R in epochs, an algorithm that we obtain by modifying Exp4.P. The "R" denotes a feature of Exp4.R: it enables inference of correct expert rankings with high probability in addition to satisfying a regret upper bound of the same order as that proved for Exp4.P. Our main result establishes a high-probability upper bound of $\tilde{\mathcal{O}}\big((i^*)^{1/\alpha}K + \sqrt{\alpha KT}\big)$ on the regret of BEES, hiding only polylog factors, which adapts to the index of the unknown best expert i^* and depends on a positive integer-valued parameter α. The experts can be ordered using domain knowledge before being input into BEES where the ones that are believed to perform well get low indices. The regret upper bound shows that domain knowledge improves the performance of BEES. The bound also illustrates the trade-off, controlled by α, between exploration and exploitation for the problem of nonstochastic bandits with infinitely many experts. On the one hand, it is desirable to include numerous experts per epoch so as to approach i^* at a fast rate. On the other hand, querying too many experts simultaneously necessitates long epochs, which reduces the rate at which more experts are included. Although tuning α needs the unknown index i^*, we can simply set $\alpha = 1$. Since we make no statistical assumptions about the rewards or the experts, the best expert i^* can depend on the time horizon T. Our regret upper bound is optimal up to polylog factors when $i^* = \tilde{\mathcal{O}}\big(\sqrt{T/K}\big)$. This regime is less restricted than it seems at first sight. If we assume a prior distribution on i^*, then $i^* = \tilde{\mathcal{O}}\big(\sqrt{T/K}\big)$ holds with a probability that increases with T, the rate of which can be fast. Inspired by the problem of finite-time model selection for reinforcement learning (RL), we also present an example of structured experts, which simulates the trade-off between approximation and estimation. We discuss how the expert ranking property of Exp4.R can be used to expedite learning in such case and

demonstrate the improvement in numerical experiments.

2.1.3 Related Work

A natural approach is to consider experts as arms and use methods for infinitely many-armed bandits such as [20, 80, 118, 32]. However, such work relies on statistical assumptions, whereas our setting is nonstochastic. Our question is also related to bandits with limited advice, first posed by [120] and subsequently solved by [75], but their results are restricted to finitely many experts. For the setting considered in this chapter, existing work either achieves a high-probability regret bound larger than $\tilde{\mathcal{O}}\big(\sqrt{KT}\big)$ or has worse computational efficiency. When configured correctly, Exp4 has a regret upper bound of $\mathcal{O}\big(\sqrt{KT}\ln i^*\big)$ *in expectation* [56]. However, the algorithm is computationally unfeasible as it needs to handle infinitely many experts at every time step. One method of making Exp4 computationally tractable is to truncate the sequence of experts to a subset of size $\mathcal{O}\big(e^{\sqrt{KT}}\big)$ as any larger set would make the expected regret superlinear in K or T. Running Exp4 with correct configurations on this subset of experts has a regret upper bound of $\mathcal{O}\big((KT)^{3/4}+T\Delta\big)$ *in expectation* where Δ is the infimum upper bound on the suboptimality gaps of the experts considered. For *stochastic* contextual bandits, Exp4.P can be used as a subroutine to achieve a high-probability regret bound of $\tilde{\mathcal{O}}\big(\sqrt{dT\ln T}\big)$ with an infinite set of experts that has a finite Vapnik–Chervonenkis dimension d [22]. Since the regret analysis of Exp4.P relies on the union bound, the algorithm does not apply to infinitely many experts in the nonstochastic setting. If we run Exp4.P on a finite subset of experts of size $\Theta\big(\delta\exp\big(\sqrt{T/(16K)}\big)\big)$, the regret is then bounded from above by $\mathcal{O}\big(K^{1/4}T^{3/4}+T\Delta\big)$ with probability at least $1-\delta$. Running Exp4.P on a subset of T experts attains a high-probability regret bound of $\tilde{\mathcal{O}}\big(\sqrt{KT}\big)$ when $i^*\leq T$. Although the worst-case regret guarantee is the same order as that provided by our algorithm for sufficiently small i^*, considering a subset of experts that is fixed in advance can lead to worse performance in practice than growing the subset adaptively, as shown in our numerical experiments. Moreover, the truncation method requires knowing T a priori, which is not necessary for BEES. Since the computational complexity of Exp4.P is linear

in the number of experts for both space and runtime, running Exp4.P on T experts becomes computationally intensive for large T.

2.1.4 Organization of the Chapter

The remainder of this chapter is organized as follows. Section 2.2 presents our main results. We first introduce Exp4.R for the setting of finitely many experts and prove that it enables inference of correct expert rankings with high probability. We then investigate the case of infinitely many experts and propose a meta-algorithm that runs Exp4.R as a subroutine. We prove a high-probability regret upper bound and give an example to illustrate how to expedite learning when working with structured experts. Section 2.3 presents simulation results that complement our theoretical findings. We conclude this chapter in Section 2.4. Finally, Section 2.5 provides proofs deferred from the previous sections.

2.2 Main Results

2.2.1 Nonstochastic Bandits with a Finite Number of Experts

We start with a simplified problem where the number of experts is finite. We first present Exp4.R (Algorithm 1) and provide some intuition for its design. We then show that Exp4.R not only preserves the regret upper bound of Exp4.P in terms of order but also enables inference of correct expert rankings with high probability.

Proposed Algorithm

Exp4.R (Algorithm 1) is a slight variant of Exp4.P proposed by [22]. The major distinction is that Exp4.R calculates a threshold vector ϵ which enables inference of correct expert rankings with high probability. Exp4.R takes four inputs, namely, an *error rate* $\delta \in (0, 1]$, a time horizon $T \in \mathbb{Z}_+$, the minimum probability $\rho \in (0, 1/K]$ of exploration, and a finite set of experts $\mathcal{I} \subset \mathbb{Z}_+$. Without loss of generality, we suppose that $|\mathcal{I}| = N$.

Exp4.R first initializes a weight $w_i(1) = 1$ for each expert $i \in \mathcal{I}$. At time $t \in [T]$, normalizing $w(t)$ gives a probability distribution $q(t)$ over $\mathcal{I}$. After getting advice $\xi^i(t)$ from each expert i, Exp4.R constructs a probability distribution $p(t)$ over $\mathcal{A}$ by weighting all advice according to $q(t)$ and mixing in uniform exploration so that $p_a(t) \geq \rho$ for all $a \in \mathcal{A}$. Specifically, for all a, let

$$p_a(t) = (1 - K\rho) \sum_{i \in \mathcal{I}} q_i(t) \xi_a^i(t) + \rho. \tag{2.1}$$

Exp4.R subsequently takes action $a(t)$ sampled according to $p(t)$ and receives the reward $r_{a(t)}(t)$. Time t concludes with weight updates as specified below. For $i \in \mathcal{I}$, Exp4.R estimates $y_i(t)$ by $\hat{y}_i(t)$ and calculates an upper bound on the variance of $\hat{y}_i(t)$ conditional on history until time $t-1$ as given by

$$\hat{y}_i(t) = \frac{\xi_{a(t)}^i(t) r_{a(t)}(t)}{p_{a(t)}(t)}, \quad \hat{v}_i(t) = \sum_{a \in \mathcal{A}} \frac{\xi_a^i(t)}{p_a(t)}. \tag{2.2}$$

Exp4.R updates each expert's weight $w_i(t)$ using

$$w_i(t+1) = w_i(t) \exp\left(\frac{\rho}{2} \left[\hat{y}_i(t) + \beta \hat{v}_i(t)\right]\right), \tag{2.3}$$

where $\beta = \sqrt{\ln(2N/\delta)/(KT)}$. The game ends in T time steps and gives two outputs, namely, the final weight vector $w(T+1)$ and a threshold vector ϵ, the ith entry of which is

$$\epsilon_i = \left[1 + \frac{1}{KT} \sum_{t=1}^{T} \hat{v}_i(t)\right] \ln\left(\frac{2N}{\delta}\right).$$

Properties

We establish in Proposition 1 that, with high probability, Exp4.R not only satisfies a regret upper bound of the same order as that proved for Exp4.P but also reveals correct pairwise expert rankings if the corresponding weights are sufficiently separated. We give some intuition here and provide proofs in Section 2.5.

For simplicity of notation, we denote $R_i([T]) \triangleq \sum_{t=1}^{T} y_i(t)$ as $R_i(T)$. Updating

Algorithm 1 Exp4.R

Input: $\delta \in (0,1]$, $T \in \mathbb{Z}_+$, $\rho \in (0, 1/K]$, $\mathcal{I} \subset \mathbb{Z}_+$

Output: $w(T+1)$, ϵ

$\beta \leftarrow \sqrt{\ln(2N/\delta)/(KT)}$.

$w_i(1) \leftarrow 1$ for $i \in \mathcal{I}$.

for $t = 1, \ldots, T$ **do**

Get $\xi^i(t)$ for $i \in \mathcal{I}$.

$q_i(t) \leftarrow w_i(t) / \sum_{i' \in \mathcal{I}} w_{i'}(t)$ for $i \in \mathcal{I}$.

$p_a(t) \leftarrow (1 - K\rho) \sum_{i \in \mathcal{I}} q_i(t) \xi_a^i(t) + \rho$ for $a \in \mathcal{A}$.

Sample action $a(t)$ from $p(t)$.

Take action $a(t)$ and receive reward $r_{a(t)}(t)$.

for $i \in \mathcal{I}$ **do**

$$\hat{y}_i(t) \leftarrow \frac{\xi^i_{a(t)}(t) r_{a(t)}(t)}{p_{a(t)}(t)},$$

$$\hat{v}_i(t) \leftarrow \sum_{a \in \mathcal{A}} \frac{\xi_a^i(t)}{p_a(t)},$$

$$w_i(t+1) \leftarrow w_i(t) \exp\left(\frac{\rho}{2}\left[\hat{y}_i(t) + \beta \hat{v}_i(t)\right]\right).$$

end for

end for

for $i \in \mathcal{I}$ **do**

$$\epsilon_i \leftarrow \left[1 + \frac{1}{KT} \sum_{t=1}^{T} \hat{v}_i(t)\right] \ln\left(\frac{2N}{\delta}\right).$$

end for

weights using (2.3) allows us to construct a confidence bound for each $R_i(T)$. For $i \in \mathcal{I}$, let $\hat{R}_i(T) \triangleq \sum_{t=1}^{T} \hat{y}_i(t)$ and $\hat{V}_i(T) \triangleq \sum_{t=1}^{T} \hat{v}_i(t)$. For any $\delta \in (0,1]$, let $\mathfrak{E}(\delta)$ be an event defined by

$$\forall i \in \mathcal{I}, \quad -\ln\left(\frac{2N}{\delta}\right)\sqrt{\frac{KT}{\ln N}} - \sqrt{\frac{\ln N}{KT}}\hat{V}_i(T) \le R_i(T) - \hat{R}_i(T) \le \sqrt{\ln\left(\frac{2N}{\delta}\right)}\left(\frac{\hat{V}_i(T)}{\sqrt{KT}} + \sqrt{KT}\right).$$

Lemma 1 shows that the estimates $\hat{R}_i(T)$ are concentrated around the true values $R_i(T)$. The proof relies on a Freedman-style inequality for martingales from [22].

Assumption 1. The following conditions hold:

(i) $\max\{4K \ln N, \ln(2N/\delta)/[(e-2)K]\} \leq T$,

(ii) and there exists a *uniform expert* $i \in \mathcal{I}$ such that $\xi_a^i(t) = 1/K$ for all $a \in \mathcal{A}$ and $t \in \mathbb{Z}_+$.

Lemma 1. *Under Assumption 1, if we run Exp4.R with* $\rho = \sqrt{\ln N/(KT)}$, *then* $\mathbb{P}(\mathfrak{E}(\delta)) \geq 1 - \delta$ *for all* $\delta \in (0, 1]$.

Lemma 2 establishes an upper bound on the regret of Exp4.R. Since Lemma 2 is a slight variant of Theorem 2 in [22], the proof is very similar to the original one and hence omitted. We note that Theorem 2 in [22] holds for a smaller regime than stated in the original paper. To be specific, the condition $T = \Omega(K \ln N)$ is essential for $\rho = \sqrt{\ln N/(KT)} \leq 1/K$ to be true. We make the correction in Lemma 2.

Lemma 2. *Under Assumption 1, for any* $\delta \in (0, 1]$, *if* $\mathfrak{E}(\delta)$ *holds, then Exp4.R with* $\rho = \sqrt{\ln N/(KT)}$ *satisfies that Regret*$(T; \mathcal{I}) \leq 7\sqrt{KT \ln(2N/\delta)}$.

Lemma 3 validates the correctness of the inferred expert rankings when the concentration event $\mathfrak{E}(\delta)$ holds. Corollary 1 shows that the uncertainty gap for ranking any pair of experts is the sum of their thresholds given by Exp4.R. We can prove Corollary 1 by first taking the contrapositive of the statement in Lemma 3 and then switching i and i'.

Lemma 3. *Under Assumption 1, for any* $\delta \in (0, 1]$, *if* $\mathfrak{E}(\delta)$ *holds, then Exp4.R with* $\rho = \sqrt{\ln N/(KT)}$ *satisfies that, for all* $i, i' \in \mathcal{I}$, *if* $\ln w_i(T+1) - \ln w_{i'}(T+1) > \epsilon_i$, *then* $R_i(T) > R_{i'}(T)$.

Corollary 1. *Under the conditions of Lemma 3, it holds that, for all* $i, i' \in \mathcal{I}$,

(i) if $\ln w_i(T+1) - \ln w_{i'}(T+1) > \epsilon_i$, *then* $R_i(T) > R_{i'}(T)$;

(ii) if $R_i(T) \geq R_{i'}(T)$, *then* $\ln w_i(T+1) - \ln w_{i'}(T+1) \geq -\epsilon_{i'}$.

Finally, we combine the lemmas to obtain Proposition 1. Same as Exp4.P, the computational complexity of Exp4.R is $\mathcal{O}(KN)$ for space and $\mathcal{O}(KNT)$ for runtime.

Proposition 1. *Under Assumption 1, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, Exp4.R configured with $\rho = \sqrt{\ln N/(KT)}$ satisfies that*

(i) Regret$(T; \mathcal{I}) \leq 7\sqrt{KT \ln(2N/\delta)}$;

(ii) for all $i, i' \in \mathcal{I}$, if $\ln w_i(T+1) - \ln w_{i'}(T+1) > \epsilon_i$, then $R_i(T) > R_{i'}(T)$.

2.2.2 Selection Among Infinitely Many Experts

In this section, we study the problem of nonstochastic bandits with a countably infinite set of experts. We make no assumptions about the experts or how they are indexed. For this general case, we propose a meta-algorithm called **Be**st **E**xpert **S**earch (BEES, Algorithm 2) that runs Exp4.R as a subroutine and provide a high-probability upper bound on regret. We also provide an example of structured experts and discuss how the expert ranking property of Exp4.R can be used to expedite learning in such case.

BEES takes five inputs including an error rate $\delta \in (0, 1]$, the number of *epochs* $L \in \mathbb{Z}_+$, and three constants $\alpha, c, C \in \mathbb{Z}_+$ that control the exponential growth of the epoch length and the number of experts queried in each epoch. At a high level, BEES supplies Exp4.R with an increasing (but still finite) number of experts over epochs, prioritizing those with lower indices. This scheme can be considered as putting a prior on the experts *implicitly* where the experts that are believed to perform well are given low indices. The regret upper bound established in Theorem 1 for BEES adapts to the unknown difficulty of the problem in the sense that i^* being large corresponds to a bad implicit prior. Since we make no assumptions about the experts, they can be ordered using domain knowledge before being input into BEES. Growing the epoch length and the number of experts at exponential rates allows us to derive a regret upper bound of the same order as that of Exp4.R when the best expert i^* has a relatively low index. This idea is similar to, though not the same as, the doubling

Algorithm 2 Best Expert Search (BEES)

Input: $\delta \in (0,1]$, $\alpha \in \mathbb{Z}_+$, $L \in \mathbb{Z}_+$, $c \in \mathbb{Z}_+$, $C \in \mathbb{Z}_+$
for epoch $l = 1, \ldots, L$ **do**
    $N_l \leftarrow c2^{\alpha l}$, $T_l \leftarrow C2^l$.
    $\rho_l \leftarrow \sqrt{\ln N_l/(KT_l)}$.
    $\mathcal{I}_l \leftarrow [N_l]$.
    Exp4.R$(\delta/L, T_l, \rho_l, \mathcal{I}_l)$.
end for

trick [21] as the latter only deals with the epoch length. We need to increase the number of experts at an appropriate rate relative to the epoch length.

Corollary 2 simplifies the bound in Theorem 1 for specific parameter values. Corollary 2 shows that BEES, when tuned right, satisfies $\text{Regret}(T) = \tilde{\mathcal{O}}\Big((i^*)^{1/\alpha}K + \sqrt{\alpha KT}\Big)$ with high probability, where $\tilde{\mathcal{O}}(\cdot)$ omits only polylog factors. This upper bound illustrates the trade-off between exploration and exploitation for the problem of bandits with infinitely many experts. On the one hand, we want to include numerous experts in each epoch so as to approach i^* fast. On the other hand, querying too many experts simultaneously necessitates long epochs, which reduces the rate at which more experts are included. This trade-off is controlled by $\alpha \in \mathbb{Z}_+$. The term $\tilde{\mathcal{O}}\left((i^*)^{1/\alpha}K\right)$ in the bound is due to not considering i^* sooner. The other term $\tilde{\mathcal{O}}\left(\sqrt{\alpha KT}\right)$ is the regret that benchmarks against the best expert in each epoch. Another consideration for not using an arbitrarily large value of α is that the minimum time horizon required by BEES which is $T = \Omega(C(\alpha, c, K, \delta))$ increases with α. Although tuning α needs the unknown index i^* of the best expert, we can simply set $\alpha = 1$. BEES has space complexity $\mathcal{O}\left(K(1+T/K)^\alpha\right)$ and time complexity $\tilde{\mathcal{O}}\left(K^2(1+T/K)^{\alpha+1}\right)$.

The regret bound in Theorem 1 matches the lower bound $\tilde{\Omega}(\sqrt{KT})$ derived by [3] up to polylog factors when $i^* = \tilde{\mathcal{O}}\big(\sqrt{T/K}\big)$. This regime is less restricted than it seems at first sight. Assuming a prior distribution on i^* shows that the condition on i^* is satisfied with a probability that increases with T, the rate of which can be fast. For simplicity, let $\alpha = 1$ and $c = 1$. In order for $\text{Regret}(T) = \tilde{\mathcal{O}}\left(\sqrt{KT}\right)$ to hold with high probability, we need $i^* = \tilde{\mathcal{O}}\big(\sqrt{T/K}\big)$. We denote the complement of this

event as $\mathfrak{B}$. If we suppose that $F(i) = \mathbb{P}(i^* > i)$ for $i \in \mathbb{Z}_+$ and some non-increasing function $F : \mathbb{Z}_+ \to [0, 1]$, then $\mathbb{P}(\mathfrak{B})$ decreases with T. For example, if $F(i) \propto i^{-s}$ for some $s > 0$, then $\mathbb{P}(\mathfrak{B})$ is roughly proportional to $K^{s/2}T^{-s/2}$. If $F(i) \propto e^{-si}$ for some $s > 0$, then $\mathbb{P}(\mathfrak{B})$ is roughly proportional to $e^{-s\sqrt{T/K}}$.

Although the worst-case regret guarantee of BEES is the same order as that achieved by running Exp4.P on a subset of T experts for sufficiently small i^*, BEES can be configured to expedite learning by exploiting the expert structure if it is known. Section 2.3 will show in numerical experiments that growing a subset of experts adaptively can improve performance in practice in comparison with fixing a subset of experts a priori. Moreover, the truncation method requires knowledge of T, which is not necessary for BEES as we can use sufficiently small δ instead of δ/L in the subroutine Exp4.R. Finally, since the computational complexity of Exp4.P is linear in the number of experts for both space and runtime, running Exp4.P on T experts becomes computationally intensive for large T.

Before stating Theorem 1, we provide some intuition for the proof. Lemma 1 implies that $\sum_{t\in\mathcal{T}_l} \hat{y}_i(t) \approx R_i(\mathcal{T}_l)$ for each expert i and every epoch l with high probability. For this reason, we can prove an upper bound on the regret with respect to the best expert in each epoch, namely, $\sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l) - \sum_{t=1}^{T} r_{a(t)}(t) = \tilde{\mathcal{O}}\left(\sqrt{\alpha KT}\right)$. We then derive an upper bound on the gap between the globally best expert and the best expert in each epoch, which is given by $R_{i^*}([T]) - \sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l) = \tilde{\mathcal{O}}\left((i^*)^{1/\alpha}K\right)$. Adding the upper bounds, we get $\text{Regret}(T) = \tilde{\mathcal{O}}\left((i^*)^{1/\alpha}K + \sqrt{\alpha KT}\right)$.

For simplicity of notation, we suppose that the total number of epochs is $L = \log_2(1 + T/(2C))$ so that $T = \sum_{l=1}^{L} T_l$ where $T_l = C2^l$ for $l \in [L]$. We use $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ to denote the floor and ceiling functions, respectively. For the general case of $T \geq 2C$, let $L = \lfloor\log_2(1 + T/(2C))\rfloor$, $T_l = C2^l$ for $l \in [L-1]$, and $T_L = T - \sum_{l=1}^{L-1} T_l$.

Theorem 1. *If a uniform expert is available in each epoch, then there exist absolute constants $\alpha \in \mathbb{Z}_+$ and $c \in \mathbb{Z}_+$ such that, for some $C(\alpha, c, K, \delta) \in \mathbb{Z}_+$, BEES satisfies*

that, for any $\delta \in (0,1]$, with probability at least $1-\delta$, we have

$$\textit{Regret}(T) < 20\sqrt{\alpha K(T+2C)\ln\left(\frac{cL(2+T/C)}{\delta}\right)} + 2C\left(\frac{i^*}{c}\right)^{\frac{1}{\alpha}}.$$

Corollary 2. *Under the conditions of Theorem 1, running BEES with $\alpha \in \mathbb{Z}_+$, $c \in \mathbb{Z}_+$, and $C = \lceil \alpha K \ln(16c^4/\delta) \rceil$ satisfies that, for any $\delta \in (0,1]$, with probability at least $1-\delta$, $\textit{Regret}(T) = \tilde{\mathcal{O}}\left((i^*)^{1/\alpha}K + \sqrt{\alpha KT}\right)$.*

Proof of Theorem 1. We can show that, for all $\delta \in (0,1]$, $\alpha \in \mathbb{Z}_+$, and $c \in \mathbb{Z}_+$, there exists $C(\alpha, c, K, \delta) \in \mathbb{Z}_+$ such that $4K\ln\left(c2^{\alpha l}\right) \le C2^l$ and $\ln\left(c2^{\alpha l+1}/\delta\right) \le (e-2)CK2^l$ for all $l \in \mathbb{Z}_+$. For example, we can set $C = \lceil \alpha K \ln(16c^4/\delta) \rceil$. Together with the definitions of N_l and T_l in Algorithm 2, we have that, for all $\alpha \in \mathbb{Z}_+$ and $c \in \mathbb{Z}_+$, there exists $C \in \mathbb{Z}_+$ such that $4K\ln N_l \le T_l$ and $\ln(2N_l/\delta) \le (e-2)KT_l$ for all $l \in \mathbb{Z}_+$. We fix such integers $\alpha, c, C \in \mathbb{Z}_+$ for the rest of the proof.

For simplicity of notation, we first consider running Exp4.R$(\delta, T_l, \rho_l, \mathcal{I}_l)$ in each epoch l for any $\delta \in (0, 1/L]$ and then apply a change of variables at the end of the proof. We suppose that a uniform expert is available in each epoch. Assumption 1 is then satisfied for all epochs. For now, we assume that event $\mathfrak{E}(\delta)$ holds for all epochs, the probability of which will be discussed at the end of the proof. For simplicity of notation, let $i_l^* \triangleq i^*(\mathcal{I}_l; \mathcal{T}_l)$ for $l \in [L]$.

Let $U_l \triangleq \alpha l + \log_2(2c/\delta)$ for $l \in [L]$. Recall that $\mathcal{T}_l$ is the time interval of epoch l where $|\mathcal{T}_l| = T_l$. By Lemma 2,

$$\begin{aligned}\sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l) - \sum_{t=1}^{T} r_{a(t)}(t) \le \sum_{l=1}^{L} 7\sqrt{KT_l\ln\left(\frac{2N_l}{\delta}\right)} &= 7\sqrt{KC\ln 2}\sum_{l=1}^{L}\sqrt{2^l U_l} \\ &\le 7\sqrt{KCU_L\ln 2}\sum_{l=1}^{L} 2^{l/2} \\ &< 20\sqrt{KCU_L}\left(2^{L/2}-1\right).\end{aligned}$$

Since $L = \log_2[1 + T/(2C)]$, we have

$$\begin{aligned}\sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l) - \sum_{t=1}^{T} r_{a(t)}(t) &< 20\sqrt{KCU_L}\left(\sqrt{1+\frac{T}{2C}}-1\right) \\ &< 20\sqrt{K\left[\alpha L + 2\ln\left(\frac{2c}{\delta}\right)\right]\left(C+\frac{T}{2}\right)}.\end{aligned} \tag{2.4}$$

We first discuss the case where $i^* \notin \mathcal{I}_1$. Let L' be the last epoch such that i^* is not considered in Algorithm 2. Since $|\mathcal{I}_l| = N_l$, we have $L' = \min\left(L, \lceil \alpha^{-1}\log_2(i^*/c)\rceil - 1\right)$. Since $i^* \notin \mathcal{I}_1$, we get $L' \geq 1$. By the definition of i_l^*, we have $R_{i_l^*}(\mathcal{T}_l) \geq R_{i^*}(\mathcal{T}_l)$ for all $l > L'$. Thus,

$$R_{i^*}([T]) - \sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l) \leq \sum_{l=1}^{L'}\left(R_{i^*}(\mathcal{T}_l) - R_{i_l^*}(\mathcal{T}_l)\right) \leq \sum_{l=1}^{L'} T_l < C2^{L'+1} < 2C\left(\frac{i^*}{c}\right)^{\frac{1}{\alpha}}. \tag{2.5}$$

We now consider the case where $i^* \in \mathcal{I}_1$. It follows from Algorithm 2 that $i^* \in \mathcal{I}_l$ for all l. Thus, the definition of i_l^* implies that $R_{i_l^*}(\mathcal{T}_l) \geq R_{i^*}(\mathcal{T}_l)$ for all l. We define $D \triangleq R_{i^*}([T]) - \sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l)$. We then have $D \leq 0$. However, the definition of i^* implies that $D \geq 0$. Therefore, $D = 0$ and (2.5) is satisfied.

Adding (2.4) and (2.5) gives

$$\text{Regret}(T) < 20\sqrt{K\left[\alpha L + 2\ln\left(\frac{2c}{\delta}\right)\right]\left(C+\frac{T}{2}\right)} + 2C\left(\frac{i^*}{c}\right)^{\frac{1}{\alpha}}. \tag{2.6}$$

Using Lemma 1 and the union bound over all L epochs, we conclude that (2.6) holds with probability at least $1 - L\delta$. A change of variables gives that, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, we have

$$\begin{aligned}\text{Regret}(T) <& 20\sqrt{K\left[\alpha L + 2\ln\left(\frac{2cL}{\delta}\right)\right]\left(C+\frac{T}{2}\right)} + 2C\left(\frac{i^*}{c}\right)^{\frac{1}{\alpha}} \\ <& 20\sqrt{\alpha K(T+2C)\ln\left(\frac{cL(2+T/C)}{\delta}\right)} + 2C\left(\frac{i^*}{c}\right)^{\frac{1}{\alpha}}.\end{aligned}$$

□

Structured Experts

We present an example of structured experts that is inspired by the problem of finite-time model selection for RL and discuss how the expert ranking property of Exp4.R can be used to expedite learning in such case.

As RL becomes increasingly integrated into autonomous systems such as agile robots [68], self-driving vehicles [82], customized fertilizer formulation [24], and personalized medication dosing [100], it is crucial that the techniques are robust [93]. An aspect of robustness is the capability to detect and adjust for model errors. For RL, this entails both model selection and parameter estimation. How to achieve both objectives simultaneously while maintaining provably good performance is an active area of research [104, 1]. The crux of the problem of online model selection for RL is to balance approximation and estimation errors in a time-dependent manner. As an example, we suppose that there is an infinite sequence of nested model classes. This structure arises naturally when an RL algorithm incorporates increasingly many features over time. Some new features may also just become obtainable while an RL algorithm is running. In fact, it is unknown a priori for many applications what is a minimal feature space that contains an optimal policy. Given an infinite sequence of model classes, the best class to use depends on the horizon or, equivalently, the amount of trajectory data that will become available. Although a larger model class has a smaller approximation error, it tends to have a higher estimation error for a fixed finite horizon. Moreover, if several classes have the same approximation power, the simplest one is typically preferred in consideration of time and space complexity.

Inspired by the problem of finite-time model selection for RL, we propose to consider experts structured in a way that simulates the trade-off between approximation and estimation. In particular, we suppose that the experts are ranked in ascending order of complexity. We propose a variant of BEES which also operates in $L = \mathcal{O}(\ln T)$ epochs with $\mathcal{T}_l$ being the time interval of epoch l. Assumption 2 stipulates that the total reward is weakly unimodal in expectation with respect to the expert index during any epoch. In addition, the index of the globally best expert is nondecreasing as

the epoch increases. See Fig. 2-1 for an illustration. Section 2.3 will demonstrate in numerical experiments that a noisy unimodal structure can be sufficient in practice.

Assumption 2. For any epoch $l \in [L]$, if $i \leq i^*(\mathbb{Z}_+; \mathcal{T}_l)$, then $R_{i-1}(\mathcal{T}_l) \leq R_i(\mathcal{T}_l)$. Otherwise, $R_i(\mathcal{T}_l) \geq R_{i+1}(\mathcal{T}_l)$. Moreover, $i^*(\mathbb{Z}_+; \mathcal{T}_l) \leq i^*(\mathbb{Z}_+; \mathcal{T}_{l'})$ if $l < l'$.

The proposed time-dependent unimodal structure is fundamentally related to oracle inequalities in empirical risk minimization [143]. Although the experts' performance may fluctuate around the proposed structure in practice, solutions to the stylized setting are of theoretical interest. Unimodal bandits have been previously studied for the stochastic setting where the expected reward is a unimodal function of partially ordered arms [43, 151, 41, 42]. Extensions to non-stationary environments have been proposed for low-frequency abrupt changes [151] and smooth changes [41] in expected rewards. Our setting is a nonstochastic bandit problem with no assumptions on the frequency or the magnitude of changes in the unimodal structure.

Under Assumption 2, the outputs of Exp4.R give a threshold rule that allows us to find a lower bound for i^*, which can accelerate the rate of approaching i^*. We modify BEES to incorporate lower bound estimation (BEES.LB, Algorithm 3). BEES.LB runs Exp4.R and Probabilistic Thresholding Search (PTS, Algorithm 4) as subroutines. In each epoch, BEES.LB eliminates experts identified as suboptimal. Lemma 4 shows that the estimated lower bound is correct if the concentration event $\mathfrak{E}(\delta)$ holds. Theorem 2 establishes a high-probability regret upper bound for BEES.LB. The proof is similar to that of Theorem 1, hence deferred until Section 2.5. PTS has space complexity $\mathcal{O}(N)$ and time complexity $\mathcal{O}(N^2)$. PTS can be efficiently implemented by first sorting the input w. BEES.LB takes the same space $\mathcal{O}(K(1+T/K)^\alpha)$ as BEES. The time complexity of BEES.LB is $\tilde{\mathcal{O}}(K^2(1+T/K)^{\alpha+1} + (1+T/K)^{2\alpha})$, which reduces to the runtime of BEES for sufficiently small α.

Lemma 4. *Under Assumption 2 and the conditions of Lemma 3, if event $\mathfrak{E}(\delta)$ holds for all epochs, then $\underline{i}_l \leq i^*$ for all l.*

Theorem 2. *Under Assumption 2, if a uniform expert is available in each epoch, then there exist absolute constants $\alpha \in \mathbb{Z}_+$ and $c \in \mathbb{Z}_+$ such that, for some $C(\alpha, c, K, \delta) \in$*

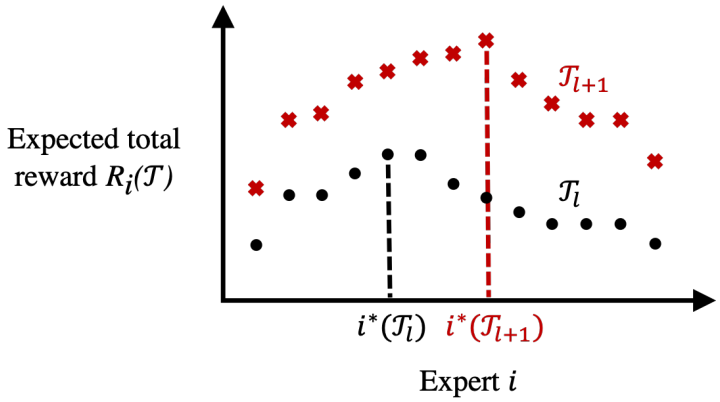

Figure 2-1: An illustration of Assumption 2.

Algorithm 3 BEES with **L**ower **B**ound (BEES.LB)

Input: $\delta \in (0, 1]$, $\alpha \in \mathbb{Z}_+$, $L \in \mathbb{Z}_+$, $c \in \mathbb{Z}_+$, $C \in \mathbb{Z}_+$
$\underline{i}_1 \leftarrow 1$.
for epoch $l = 1, \ldots, L$ **do**
    $N_l \leftarrow c2^{\alpha l}$, $T_l \leftarrow C2^l$.
    $\rho_l \leftarrow \sqrt{\ln N_l / (KT_l)}$.
    $\mathcal{I}_l \leftarrow \{\underline{i}_l, \underline{i}_l + 1, \ldots, \underline{i}_l + N_l - 1\}$.
    $w^l, \epsilon^l \leftarrow \text{Exp4.R}(\delta/L, T_l, \rho_l, \mathcal{I}_l)$.
    $\underline{i}_{l+1} \leftarrow \text{PTS}\left(w^l, \epsilon^l, \underline{i}_l\right)$.
end for

Algorithm 4 Probabilistic Thresholding Search (PTS)

Input: $w \in (0, \infty)^N$, $\epsilon \in (0, \infty)^N$, $\underline{i} \in \mathbb{Z}_+$
Output: $\underline{i}_{\text{new}}$
$\underline{j} \leftarrow 1$.
for $j = 1, \ldots, N - 1$ **do**
    for $j' = j + 1, \ldots, N$ **do**
        if $\ln w_{j'} - \ln w_j > \epsilon_{j'}$ **then**
            $\underline{j} \leftarrow j + 1$.
        end if
    end for
end for
$\underline{i}_{\text{new}} \leftarrow \underline{i} + \underline{j} - 1$.

$\mathbb{Z}_+$, *BEES.LB satisfies that, for any* $\delta \in (0, 1]$, *with probability at least* $1 - \delta$, *we have*

$$Regret(T) < 20\sqrt{\alpha K(T + 2C) \ln\left(\frac{cL(2 + T/C)}{\delta}\right)} + 2C\left(\frac{i^*}{c}\right)^{\frac{1}{\alpha}}.$$

The upper bound in Theorem 2 is the same as that for the general case of unstructured experts because the lower bound from PTS can stay at 1 in the worst case. A trivial example is that all experts are the same. For cases where the experts' performance differs by sufficient margins, the actual improvement of BEES.LB over BEES should become obvious as we will demonstrate in Section 2.3.

If the globally best expert i^* is fixed over time, then we can modify BEES.LB to additionally estimate an upper bound on i^*, initialized to ∞. The modified search subroutine can be considered as a probabilistic counterpart of search algorithms such as the golden-section search [79]. The major difference is that the search subroutine applies to problems where the function cannot be evaluated directly. We can show that the confidence interval for i^* contracts over epochs. While the epoch length always grows exponentially, the set of experts considered in each epoch is data-dependent. If no upper bound on i^* has been identified, then the number of experts considered will increase by a factor of 2^α in the next epoch. Otherwise, only the experts in the non-expanding confidence interval will be considered from now on.

2.3 Experiments

In this section, we present two simulation results that complement our theoretical findings. The first experiment uses synthetic data to show the advantage of our proposed algorithm in the context of structured experts. In the second experiment, we apply our algorithm to crop management in digital agriculture.

We conduct numerical simulations to demonstrate the performance improvement of BEES.LB in comparison with BEES and Exp4.P when experts are structured. We consider $K = 10$ actions the rewards of which are binary and nonstochastic. The

sequence of experts has a weakly unimodal structure that is corrupted with random noise. The first expert is uniform and the best expert has index $i^* = 9$. At each time, every expert's advice is distorted with an additive K-vector that consists of independent zero-mean Gaussian noises with standard deviation 0.01, which may alter the unimodal structure of the experts. For BEES and BEES.LB, we set $\alpha = c = 1$ and C as defined in Corollary 2. We implement the version of BEES and BEES.LB that does not know the number of epochs L or the time horizon T a priori by using δ instead of δ/L in the subroutine Exp4.R. In contrast, we configure the benchmark algorithm Exp4.P with the correct T. Exp4.P is run on the first T experts in the sequence. All algorithms use an error rate $\delta = 0.05$.

Fig. 2-2 shows that BEES.LB indeed has a lower regret than BEES because of the expedited learning enabled by Exp4.R. For large T, Exp4.P is surpassed by BEES.LB as querying too many experts can increase the chance of getting bad advice. Fig. 2-2 demonstrates the advantage of our algorithm in having improved performance by being able to exploit structural information and query experts adaptively.

In the second experiment, we apply Exp4.R to learn crop management. To be specific, actions correspond to the frequencies of irrigation and fertilization. Rewards are the normalized total weight of storage organs. Fig. 2-3 illustrates our pipeline of data-driven decision making in digital agriculture. We use the **wo**rld **fo**od **st**udies (WOFOST) model [136] to simulate crop yield dependent on different management strategies. WOFOST has been an important model for crop monitoring and yield prediction in Europe for decades [46]. Although the algorithm only gets to observe the rewards of the actions taken, we can design experts using WOFOST. For example, a good expert often recommends actions that have high rewards, and a bad expert tends to suggest actions that give low rewards. The algorithm starts clueless about the experts and learns their performance through trial and error. Fig. 2-4 shows that good experts indeed tend to get consulted more over time. Fig. 2-5 corroborates Proposition 1 that Exp4.R tends to rank experts with big weight gaps correctly.

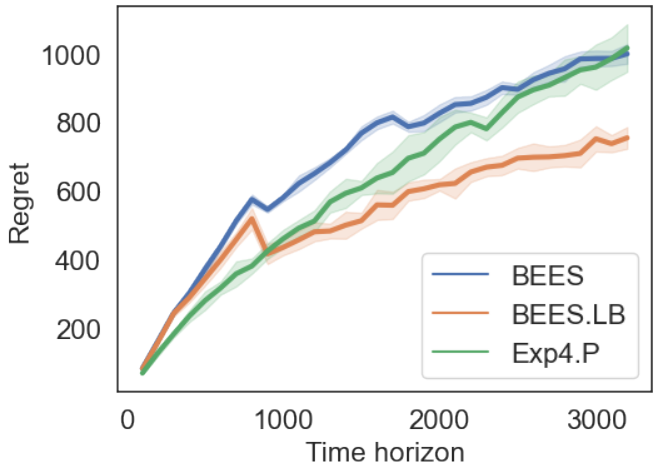

Figure 2-2: Comparison of BEES, BEES.LB, and Exp4.P in terms of regret as the time horizon varies. BEES.LB surpasses BEES and Exp4.P as the time horizon increases. Lines and shades are the averages and the standard deviations of 10 runs, respectively.

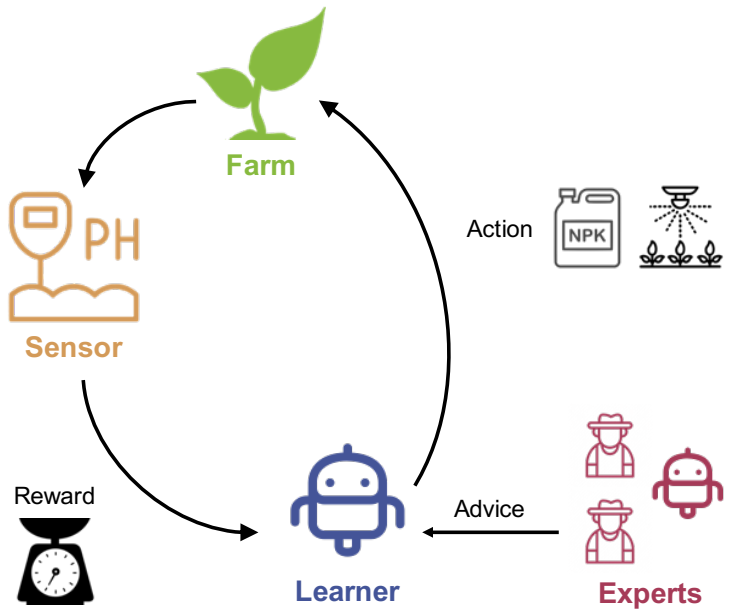

Figure 2-3: Illustration of data-driven decision making in digital agriculture.

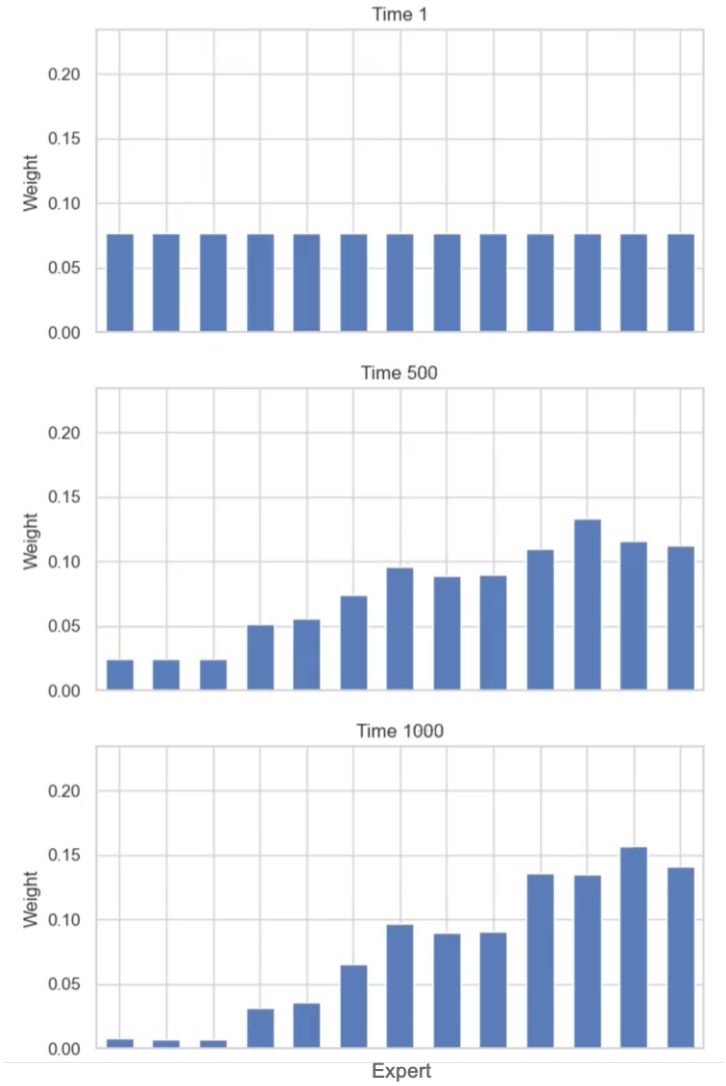

Figure 2-4: Exp4.R tends to sample good experts more often over time.

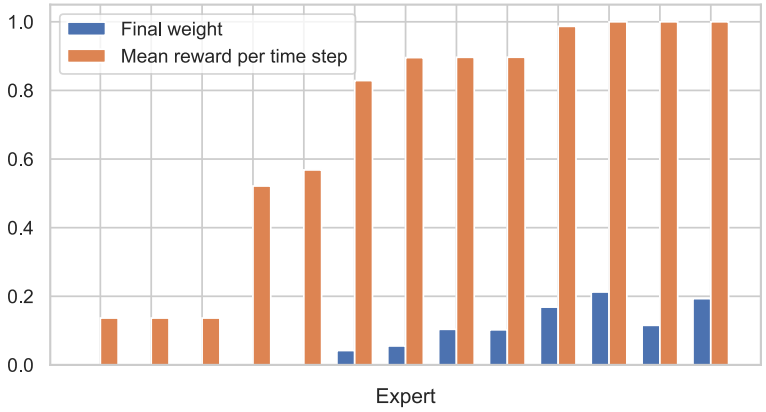

Figure 2-5: Exp4.R tends to rank experts with big weight gaps correctly.

2.4 Discussion

In this chapter, we have proposed an algorithm for the problem of nonstochastic bandits with infinitely many experts under the constraint of having access to only a finite subset of experts. We have established a high-probability upper bound on the regret of our meta-algorithm BEES, which matches the lower bound up to polylog factors if the globally best expert has a relatively low index. If we assume that there exists a prior distribution on the best expert, then the probability that our regret upper bound is tight will increase with the time horizon, the rate of which can be fast. The expert ranking property of the subroutine Exp4.R enables learning acceleration if the structure of the experts is known. We have illustrated this point with an example that is inspired by the problem of finite-time model selection for RL. One interesting direction for future work is to obtain instance-dependent upper bounds in terms of the experts' suboptimality gaps. Such instance-dependent bounds can be used to prove the learning acceleration enabled by Exp4.R. It is also worthwhile to design efficient implementation for specific applications.

2.5 Proofs

Proof of Proposition 1

Let $\mathbb{E}_t[\cdot]$ denote the conditional expectation given history until time $t-1$. We can show that $\hat{y}_i(t)$ is a conditionally unbiased estimator for $y_i(t)$. In other words, $\mathbb{E}_t[\hat{y}_i(t)] = y_i(t)$ for all i and t. Lemma 5 shows that $\hat{v}_i(t)$ is an upper bound on the conditional variance of $\hat{y}_i(t)$. Lemma 6 is a Freedman-style inequality for martingales from [22]. The proof of Lemma 1 relies on Lemmas 5 and 6.

Lemma 5 (From proof of Lemma 3 in [22]). *For all $t \in \mathbb{Z}_+$ and $i \in \mathcal{I}$, we have $\mathbb{E}_t\left[(y_i(t) - \hat{y}_i(t))^2\right] \leq \hat{v}_i(t)$.*

Lemma 6 ([22], Theorem 1). *Let $X_1, \ldots, X_T$ be a sequence of real-valued random variables. For any real-valued random variable Y, we define $\mathbb{E}_t[Y] \triangleq \mathbb{E}[Y \mid X_1, \ldots, X_{t-1}]$.*

We assume that, $X_t \leq B$ *and* $\mathbb{E}_t[X_t] = 0$ *for all* t*. We define the random variables*

$$S \triangleq \sum_{t=1}^{T} X_t, \quad V \triangleq \sum_{t=1}^{T} \mathbb{E}_t\left[X_t^2\right].$$

For any fixed estimate $V' > 0$ *of* V*, and for any* $\delta \in (0,1]$*, with probability at least* $1-\delta$*, we have*

$$S \leq \begin{cases} \sqrt{(e-2)\ln\left(\frac{1}{\delta}\right)}\left(\frac{V}{\sqrt{V'}} + \sqrt{V'}\right), & \text{if } V' \geq \frac{B^2\ln(1/\delta)}{e-2}, \\ B\ln(1/\delta) + (e-2)\frac{V}{B}, & \text{otherwise.} \end{cases}$$

Proof of Lemma 1. We now fix any $i \in \mathcal{I}$ and $t \in \mathbb{Z}_+$. By definition, we have $y_i(t) \in [0,1]$. Using (2.1) and the assumption that $\rho \in [0, 1/K]$, we get $p_a(t) \geq \rho$ for all $a \in \mathcal{A}$. Thus, (2.2) implies that $\hat{y}_i(t) \in [0, 1/\rho]$ almost surely. Let $X_t = y_i(t) - \hat{y}_i(t)$. We then have $-1/\rho \leq X_t \leq 1$ almost surely. We can show that $\mathbb{E}_t[\hat{y}_i(t)] = y_i(t)$ and hence $\mathbb{E}_t[X_t] = 0$. We recall that $R_i(T) = \sum_{t=1}^{T} y_i(t)$. Applying Lemma 6 to $(X_t)_t$ and $(-X_t)_t$ respectively and then taking a union bound, we conclude that, for any $\delta \in (0,1]$, with probability at least $1-\delta/N$, the inequality $-B_1 \leq R_i(T) - \hat{R}_i(T) \leq B_2$ holds, where

$$B_1 \triangleq \begin{cases} \sqrt{(e-2)\ln\left(\frac{2N}{\delta}\right)}\left(\frac{V}{\sqrt{V'}} + \sqrt{V'}\right), & \text{if } V' \geq \frac{\ln(2N/\delta)}{(e-2)\rho^2}, \\ \frac{\ln(2N/\delta)}{\rho} + (e-2)\rho V, & \text{otherwise,} \end{cases}$$

$$B_2 \triangleq \begin{cases} \sqrt{(e-2)\ln\left(\frac{2N}{\delta}\right)}\left(\frac{V}{\sqrt{V'}} + \sqrt{V'}\right), & \text{if } V' \geq \frac{\ln(2N/\delta)}{(e-2)}, \\ \ln(2N/\delta) + (e-2)V, & \text{otherwise,} \end{cases}$$

$$V \triangleq \sum_{t=1}^{T} \mathbb{E}_t\left[X_t^2\right].$$

We now fix an arbitrary $\delta \in (0,1]$. Assumption 1 implies that $\ln(2N/\delta) \leq (e-2)KT$. Taking $\rho = \sqrt{\ln N/(KT)}$ and $V' = KT$, we have $\ln(2N/\delta)/(e-2) \leq V' < \ln(2N/\delta)/[(e-2)\rho^2]$. Lemma 5 implies that $V \leq \hat{V}_i(T)$. Therefore, with probability

at least $1-\delta/N$, we have

$$-\ln\left(\frac{2N}{\delta}\right)\sqrt{\frac{KT}{\ln N}}-\sqrt{\frac{\ln N}{KT}}\hat{V}_i(T)\le R_i(T)-\hat{R}_i(T)\le\sqrt{\ln\left(\frac{2N}{\delta}\right)}\left(\frac{\hat{V}_i(T)}{\sqrt{KT}}+\sqrt{KT}\right).$$

Applying the union bound over $i\in\mathcal{I}$, we conclude that $\mathbb{P}\left(\mathfrak{E}(\delta)\right)\ge 1-\delta$. □

Proof of Lemma 3. We fix an arbitrary $\delta\in(0,1]$ and suppose that event $\mathfrak{E}(\delta)$ holds. We recall that $\epsilon_i \triangleq \left[1+\hat{V}_i(T)/(KT)\right]\ln(2N/\delta)$. We assume that $\ln w_i(T+1)-\ln w_{i'}(T+1)>\epsilon_i$ for some $i,i'\in\mathcal{I}$. By (2.3) and the initialization condition $w_i(1)=1$, we have

$$\ln w_i(T+1)=\sum_{t=1}^{T}\ln\left(\frac{w_i(t+1)}{w_i(t)}\right)=\frac{\rho}{2}\left(\hat{R}_i(T)+\sqrt{\frac{\ln(2N/\delta)}{KT}}\hat{V}_i(T)\right).$$

Thus,

$$\hat{R}_i(T)=\frac{2}{\rho}\ln w_i(T+1)-\sqrt{\frac{\ln(2N/\delta)}{KT}}\hat{V}_i(T). \tag{2.7}$$

Equation (2.7) also holds for i'. Thus,

$$\begin{aligned}
&\hat{R}_i(T)-\hat{R}_{i'}(T)\\
&=\frac{2}{\rho}\ln\left(\frac{w_i(T+1)}{w_{i'}(T+1)}\right)-\sqrt{\frac{\ln(2N/\delta)}{KT}}\left(\hat{V}_i(T)-\hat{V}_{i'}(T)\right)\\
&>\frac{2\epsilon_i}{\rho}-\sqrt{\frac{\ln(2N/\delta)}{KT}}\left(\hat{V}_i(T)-\hat{V}_{i'}(T)\right)\\
&=2\ln\left(\frac{2N}{\delta}\right)\sqrt{\frac{KT}{\ln N}}+\sqrt{\frac{\ln(2N/\delta)}{KT}}\hat{V}_{i'}(T)+\hat{V}_i(T)\sqrt{\frac{\ln(2N/\delta)}{KT}}\left[2\sqrt{\frac{\ln(2N/\delta)}{\ln N}}-1\right]\\
&>2\ln\left(\frac{2N}{\delta}\right)\sqrt{\frac{KT}{\ln N}}+\sqrt{\frac{\ln(2N/\delta)}{KT}}\left(\hat{V}_i(T)+\hat{V}_{i'}(T)\right).
\end{aligned} \tag{2.8}$$

Event $\mathfrak{E}(\delta)$ implies that

$$\begin{aligned}
R_i(T)-\hat{R}_i(T)+\hat{R}_{i'}(T)-R_{i'}(T)\ge&-\ln\left(\frac{2N}{\delta}\right)\sqrt{\frac{KT}{\ln N}}-\sqrt{\frac{\ln N}{KT}}\hat{V}_i(T)\\
&-\sqrt{\ln\left(\frac{2N}{\delta}\right)}\left(\frac{\hat{V}_{i'}(T)}{\sqrt{KT}}+\sqrt{KT}\right).
\end{aligned} \tag{2.9}$$

Adding (2.8) and (2.9) and then simplifying the algebra give

$$R_i(T) - R_{i'}(T) > 0.$$

□

Proof of Proposition 1. Proposition 1 follows directly from Lemmas 1–3. □

Proof of Theorem 2

Proof of Lemma 4. Under the assumption that event $\mathfrak{E}(\delta)$ holds for all epochs, we prove the statement by induction on l. The base case holds trivially as $\underline{i}_1 = 1$. For the inductive step, we assume that $\underline{i}_\iota \leq i^*$ for all $\iota \leq l$. If $\underline{i}_{l+1} = \underline{i}_l$, then $i^* \geq \underline{i}_{l+1}$ by the induction hypothesis. If there exists some $j \geq 1$ such that $\underline{i}_{l+1} = \underline{i}_l + j$, then Algorithm 4 implies that $\ln w_{j'} - \ln w_j > \epsilon_{j'}$ for some $j' > j$ in epoch l. Using Assumption 2 and Lemma 3, we get $i^* \geq \underline{i}_l + j = \underline{i}_{l+1}$. □

Proof of Theorem 2. We can show that, for all $\delta \in (0, 1]$, $\alpha \in \mathbb{Z}_+$, and $c \in \mathbb{Z}_+$, there exists $C(\alpha, c, K, \delta) \in \mathbb{Z}_+$ such that $4K \ln\left(c2^{\alpha l}\right) \leq C2^l$ and $\ln\left(c2^{\alpha l+1}/\delta\right) \leq (e-2)CK2^l$ for all $l \in \mathbb{Z}_+$. For example, we can set $C = \lceil \alpha K \ln(16c^4/\delta) \rceil$. Together with the definitions of N_l and T_l in Algorithm 3, we have that, for all $\alpha \in \mathbb{Z}_+$ and $c \in \mathbb{Z}_+$, there exists $C \in \mathbb{Z}_+$ such that $4K \ln N_l \leq T_l$ and $\ln(2N_l/\delta) \leq (e-2)KT_l$ for all $l \in \mathbb{Z}_+$. We fix such integers $\alpha, c, C \in \mathbb{Z}_+$ for the rest of the proof.

For simplicity of notation, we first consider running Exp4.R$(\delta, T_l, \rho_l, \mathcal{I}_l)$ in each epoch l of Algorithm 3 for any $\delta \in (0, 1/L]$ and then apply a change of variables at the end of the proof. We suppose that a uniform expert is available in each epoch. Assumption 1 is then satisfied for all epochs. For now, we assume that event $\mathfrak{E}(\delta)$ holds for all epochs, the probability of which will be discussed at the end of the proof. For simplicity of notation, let $i_l^* \triangleq i^*(\mathcal{I}_l; \mathcal{T}_l)$ for $l \in [L]$.

Let $U_l \triangleq \alpha l + \log_2(2c/\delta)$ for $l \in [L]$. Recall that $\mathcal{T}_l$ is the time interval of epoch l

where $|\mathcal{T}_l| = T_l$. By Lemma 2,

$$\begin{aligned}
\sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l) - \sum_{t=1}^{T} r_{a(t)}(t) &\leq \sum_{l=1}^{L} 7\sqrt{KT_l \ln\left(\frac{2N_l}{\delta}\right)} \\
&= \sum_{l=1}^{L} 7\sqrt{KC2^l \ln\left(\frac{c2^{\alpha l+1}}{\delta}\right)} \\
&= 7\sqrt{KC\ln 2}\sum_{l=1}^{L}\sqrt{2^l U_l} \\
&\leq 7\sqrt{KCU_L \ln 2}\sum_{l=1}^{L} 2^{l/2} \\
&< 20\sqrt{KCU_L}\left(2^{L/2}-1\right).
\end{aligned}$$

Since $L = \log_2[1 + T/(2C)]$, we have

$$\begin{aligned}
\sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l) - \sum_{t=1}^{T} r_{a(t)}(t) &< 20\sqrt{KCU_L}\left(\sqrt{1+\frac{T}{2C}}-1\right) \\
&< 20\sqrt{K\left[\alpha L + 2\ln\left(\frac{2c}{\delta}\right)\right]\left(C+\frac{T}{2}\right)}.
\end{aligned} \tag{2.10}$$

We first discuss the case where $i^* \notin \mathcal{I}_1$. Let L'' be the last epoch such that i^* is not considered in Algorithm 3. In other words, $L'' \triangleq \max\{\, l \in [L] \mid i^* \notin \mathcal{I}_l \,\}$. Lemma 4 implies that $i^* \in \mathcal{I}_l$ for all $l > L''$. By the definition of i_l^*, we have $R_{i_l^*}(\mathcal{T}_l) \geq R_{i^*}(\mathcal{T}_l)$ for all $l > L''$. Thus,

$$R_{i^*}([T]) - \sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l) \leq \sum_{l=1}^{L''}\left(R_{i^*}(\mathcal{T}_l) - R_{i_l^*}(\mathcal{T}_l)\right) \leq \sum_{l=1}^{L''} T_l < C2^{L''+1}.$$

We now provide an upper bound on L''. By Algorithms 3 and 4, we have $|\mathcal{I}_l| = N_l$ and $1 \leq \underline{i}_l \leq \underline{i}_{l+1}$ for all l. Let L' be the last epoch such that i^* is not considered in the worst case where $\underline{i}_l = 1$ for all l. In other words, $L' \triangleq \min\left(L, \lceil \alpha^{-1}\log_2(i^*/c)\rceil - 1\right)$. Under the assumption that $i^* \notin \mathcal{I}_1$, we get $L' \geq 1$. By the definitions of L' and L'',

we have $L'' \leq L'$ and hence

$$R_{i^*}([T]) - \sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l) < C2^{L'+1} < 2C\left(\frac{i^*}{c}\right)^{\frac{1}{\alpha}}. \tag{2.11}$$

We now consider the case where $i^* \in \mathcal{I}_1$. It follows from Lemma 4 that $i^* \in \mathcal{I}_l$ for all l. Thus, the definition of i_l^* implies that $R_{i_l^*}(\mathcal{T}_l) \geq R_{i^*}(\mathcal{T}_l)$ for all l. We define $D \triangleq R_{i^*}([T]) - \sum_{l=1}^{L} R_{i_l^*}(\mathcal{T}_l)$. We then have $D \leq 0$. However, the definition of i^* implies that $D \geq 0$. Therefore, $D = 0$ and (2.11) is satisfied.

Adding (2.10) and (2.11) gives

$$\text{Regret}(T) < 20\sqrt{K\left[\alpha L + 2\ln\left(\frac{2c}{\delta}\right)\right]\left(C + \frac{T}{2}\right)} + 2C\left(\frac{i^*}{c}\right)^{\frac{1}{\alpha}}. \tag{2.12}$$

Using Lemma 1 and the union bound over all L epochs, we conclude that (2.12) holds with probability at least $1 - L\delta$. A change of variables gives that, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, we have

$$\begin{aligned}\text{Regret}(T) <& 20\sqrt{K\left[\alpha L + 2\ln\left(\frac{2cL}{\delta}\right)\right]\left(C + \frac{T}{2}\right)} + 2C\left(\frac{i^*}{c}\right)^{\frac{1}{\alpha}} \\ <& 20\sqrt{\alpha K(T + 2C)\ln\left(\frac{cL(2 + T/C)}{\delta}\right)} + 2C\left(\frac{i^*}{c}\right)^{\frac{1}{\alpha}}.\end{aligned}$$

□

Chapter 3

Active Learning for Efficient Cell Reprogramming

3.1 Introduction

Cell reprogramming, the process of converting one cell type into another, has far-reaching implications for human disease modeling [97, 106], regenerative medicine [88, 110, 123], and drug screening [14, 63]. Cells can be reprogrammed in various ways including nuclear transplantation [70], cell fusion [150], and transcription-factor transduction [65]. Experiments have successfully demonstrated that overexpressing certain transcription factors (TFs) is sufficient to reprogram one particular cell type into another [30, 62, 69, 108, 124, 125, 141, 147]. We are interested in finding the combination of TFs that can make embryonic stem cells differentiate into hematopoietic stem cells (HSCs). The capability to drive directed differentiation will provide a scalable source for HSCs which are important for treating certain types of cancer and immune system disorders [17, 44].

Finding optimal interventions for gene regulatory networks is challenging because of the high-dimensional state and action spaces. Although it has been successfully demonstrated in experiments that cells can be reprogrammed by transduction of TFs, these discoveries were based on exhaustive testing of plausible TF combinations. Brute-force search is unscalable as perturbation experiments are time-consuming,

labor-intensive, and expensive. For example, there are approximately one trillion possible combinations of four human TFs, and it is impossible to test all possibilities.

Computational approaches can help improve the efficiency of cell reprogramming protocols [5, 31, 45, 48, 76, 99, 111, 116]. Existing methods typically construct some types of networks from numerous datasets and then rank TFs according to certain scores. For example, CellNet [31, 99] infers gene regulatory networks using 3419 gene expression profiles of various tissues and cell types, which enables the identification of high-influence TFs. Similarly, Mogrify [111] combines differential expression analysis and network information based on 700 libraries of gene expression data and two databases. Such a network-based approach requires a large amount of data and does not take the *sequential* design of experiments take into consideration. These challenges naturally lead to the question if computational approaches can help optimize perturbations under the constraint of insufficient data except for a few batches of experiments. We answer the question affirmatively in this chapter.

We propose an active learning framework that directly optimizes over combinations of TFs in a few batches of experiments. Although the complexity of this problem prohibits theoretical guarantees on the entire procedure, we propose a principled approach that builds upon multi-armed bandit algorithms. We demonstrate the success of our approach on gene expression data.

3.2 Problem Formulation

We introduce the mathematical formulation of single-cell perturbation in Section 3.2.1 and explain how batched experiments work in Section 3.2.2. We define the active learning problem in Section 3.2.3.

3.2.1 Single-Cell Perturbation

Let $\mathbb{Z}_+$ be the set of strictly positive integers. For $n \in \mathbb{Z}_+$, we define $[n] \triangleq \{1, 2, \ldots, n\}$. Let $p \in \mathbb{Z}_+$ be the number of genes. Let $X \in \mathbb{R}^p$ be the gene expression level of a source cell where X_i corresponds to gene i. Similarly, let $Y^{\text{target}} \in \mathbb{R}^p$ be

the gene expression level of a target cell. We know the gene expression distribution $\mathcal{P}$ of the source cell type (i.e., $X \sim \mathcal{P}$) and that of the target cell type, denoted by $\mathcal{Q}$ (i.e., $Y^{\text{target}} \sim \mathcal{Q}$). Let $\mathcal{I} \subset [p]$ be the set of TF genes that we can perturb. We describe a *perturbation* as $A \in \{0,1\}^p$ such that $A_i = 0$ whenever $i \notin \mathcal{I}$. Specifically, $A_i = 1$ if and only if TFs that overexpress gene i are *added*. Let $Y \triangleq f(X, A)$ be the gene expression level of the perturbed cell for some unknown function $f : \mathbb{R}^p \times \{0,1\}^p \to \mathbb{R}^p$. For intervention $A = a$, denote $\mathcal{P}^a$ as the interventional distribution, namely, $Y = f(X, a) \sim \mathcal{P}^a$.

Given a sparsity parameter $S \in \mathbb{Z}_+$ and some distance function d, our goal is to solve the following optimization problem:

$$\begin{aligned} \min_{a} \quad & d(\mathcal{P}^a, \mathcal{Q}) \\ \text{s.t.} \quad & \|a\|_0 \leq S \\ & a_i = 0 \qquad \forall i \notin \mathcal{I}. \end{aligned} \tag{3.1}$$

We call a feasible solution to (3.1) an *S-sparse perturbation* and denote the feasible set by $\mathcal{F}$. Specifically, our goal is to match the direction of the mean gene expression of the target cell type. In other words, $d(\mathcal{P}^a, \mathcal{Q}) \triangleq \left(1 - s\left(\mathbb{E}_{\mathcal{P}^a}\left[f(X,a)\right], \mathbb{E}_{\mathcal{Q}}\left[Y^{\text{target}}\right]\right)\right)/2$ where s is the cosine similarity function.

3.2.2 Batched Experiments

Let $T \in \mathbb{Z}_+$ be the number of batched experiments. Let $N \in \mathbb{Z}_+$ be the number of cells that can be perturbed per batch. We consider a sequencing budget $K \in \mathbb{Z}_+$. In other words, we can measure the gene expression of K cells by single-cell RNA sequencing. We note that $T \ll K \ll N$. Let $c : \{0,1\}^p \to [0,1]$ be a function that indicates lethality of a perturbation. To be specific, perturbation a kills a cell with probability $c(a)$. Let $b : \mathbb{R}^p \to \{0,1\}$ be a binary filter that picks out cells which show the partial success of reprogramming with label 1. In experiment, the filter corresponds to an enrichment step that identifies marker proteins for the target cell type. In addition to estimating $d(\mathcal{P}^a, \mathcal{Q})$ using paired information of perturbation and gene expression, we

can improve the results with enrichment outcomes. To be specific, perturbations that move cells close to the target tend to generate enriched cells (i.e, $b = 1$) as illustrated in Fig. 3-1. For our target cell type being HSC, the enrichment step uses CD34 as the marker gene. $CD34^+$ cells are essential for human hematopoiesis and are routinely used clinically as a source for bone marrow transplantation in humans [107].

We describe batched experiments as follows: for each batch $t = 1, \ldots, T$,

i (Perturbation decision) Decide on a subset of lentiviruses and their concentrations as described by $\lambda(t) \in \mathbb{R}^p$. Specifically, $\lambda_i(t)$ is the multiplicity of infection (MOI) of lentiviruses that carry gene i, where $\lambda_i(t) = 0$ if and only if we do not add lentiviruses that carry gene i. Since we only perturb TF genes, we call $\lambda \in \mathbb{R}^p$ an *MOI vector* if $\lambda_i \geq 0$ for all $i \in \mathcal{I}$ and $\lambda_i = 0$ otherwise.

ii (Viral infection) Add $\lambda(t)$ to a plate of N source cells the gene expression of which can be considered as independent and identically distributed random variables $X^1(t), \ldots, X^N(t)$ drawn from the distribution $\mathcal{P}$. For $n \in [N]$, let $A^n(t) \in \{0, 1\}^p$ be the indicator vector of the TFs that cell n has received. We can show that the average fraction of cells which have received lentiviruses that overexpress gene $i \in [p]$ is equal to $\mathbb{E}\left[N^{-1} \sum_{n=1}^{N} A_i^n(t)\right] = 1 - e^{-\lambda_i(t)}$.

iii (Cell death) Trigger cell death by sampling from Bernoulli distributions with success probabilities $\{c(A^n(t))\}_{n \in [N]}$. Let $\mathcal{S}(t) \subseteq [N]$ denote the subset of cells that have survived.

iv (Transcription) Generate $\tilde{Y}^n(t) \triangleq f_{\epsilon^n(t)}(X^n(t), A^n(t))$ for $n \in \mathcal{S}(t)$ where $f_{\epsilon^n(t)}$ is a noisy version of f that is corrupted by a random measurement error $\epsilon^n(t) \in \mathbb{R}^p$.

(We note that $\mathcal{S}(t)$ and $\{A^n(t), \tilde{Y}^n(t)\}_{n \in \mathcal{S}(t)}$ are hidden so far.)

v (Sequencing) Sample uniformly at random without replacement from $\mathcal{S}(t)$ and call the subset sampled $\mathcal{R}(t)$. Further sample uniformly at random without replacement from $CD34^+$ cells and call the subset $\mathcal{N}(t)$. We note that $\mathcal{R}(t) \cap \mathcal{N}(t) = \emptyset$ and $b\left(\tilde{Y}^n(t)\right) = 1$ for all $n \in \mathcal{N}(t)$. Fig. 3-2 illustrates the differential sampling process. Measure $A^n(t)$ and $\tilde{Y}^n(t)$ for $n \in \mathcal{N}(t) \cup \mathcal{R}(t)$.

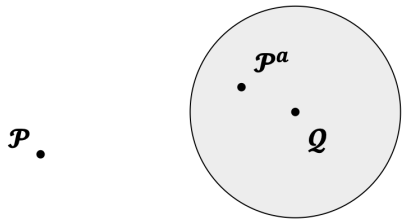

Figure 3-1: Enrichment helps optimize perturbations by identifying cells that are close to the target cell type. The gene expression distributions of the source cell type and the target cell type are denoted by $\mathcal{P}$ and $\mathcal{Q}$, respectively. The interventional distribution of perturbation a is $\mathcal{P}^a$. If $d(\mathcal{P}^a, \mathcal{Q})$ is sufficiently small, then $b(Y) = 1$ with high probability for $Y \sim \mathcal{P}^a$.

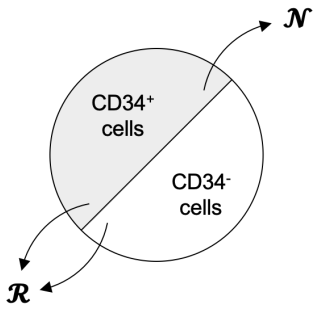

Figure 3-2: Differential sampling of enriched cells. An unbiased subset $\mathcal{R}$ is first sampled uniformly at random without replacement from all surviving cells. Among the remaining cells, a subset $\mathcal{N}$ is then sampled uniformly at random without replacement from CD34^+ cells.

Fact. If we define MOI as the ratio of the number of virions to the number of cells that are *present* in a cell-culture dish, then the probability that a cell will *receive* exactly $m \in \{0\} \cup \mathbb{Z}_+$ virions when inoculated with an MOI of $\lambda_0 > 0$ can be modeled as a Poisson distribution with mean λ_0. Thus, the average fraction of cells in a population that are each infected by exactly m virions is $\lambda_0^m e^{-\lambda_0}/(m!)$. We assume that the Poisson distributions of different lentiviruses are mutually independent. For any perturbation a, the probability that a cell is perturbed by exactly a is given by

$$\eta(a, \lambda) \triangleq \prod_{i:\, a_i = 1} \left(1 - e^{-\lambda_i}\right) \prod_{i':\, a_{i'} = 0} e^{-\lambda_{i'}}. \tag{3.2}$$

Therefore, the average fraction of cells that are perturbed by exactly a is $\eta(a, \lambda)$.

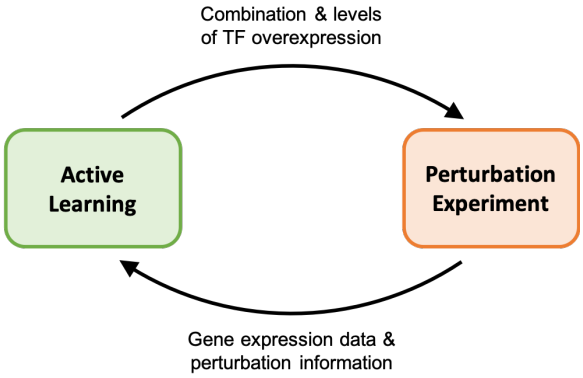

Figure 3-3: Schematic diagram of active learning and perturbation experiment integration.

3.2.3 Goal

We address in this chapter the question if it is possible to use active learning for choosing perturbation MOIs (step i in Section 3.2.2) so that we find a close-to-optimal solution to (3.1) at the end of the batched experiments. Our answer is affirmative.

In order to state the question precisely, we introduce some definitions. For notational simplicity, we assume that (3.1) has a unique solution a^*. Let $\Delta_a \triangleq d(\mathcal{P}^a, \mathcal{Q}) - d(\mathcal{P}^{a^*}, \mathcal{Q})$ be the suboptimality gap of perturbation a. Let A be the perturbation recommended by the active learning algorithm at the end of the batched experiments. We then have

$$\mathbb{E}\left[\Delta_A\right] = \sum_{a \neq a^*} \mathbb{P}\left(A = a\right) \Delta_a \leq \mathbb{P}\left(A \neq a^*\right), \tag{3.3}$$

where the last step follows from the fact that the normalized cosine distance has a bounded codomain $[0, 1]$. Equation (3.3) shows that minimizing the probability of misidentifying a^* helps finding close-to-optimal perturbations. For this reason, we will focus on minimizing $\mathbb{P}\left(A \neq a^*\right)$ under the constraint of T batched experiments. This setting is known as pure exploration with a fixed budget in multi-armed bandit problems. Fig. 3-3 summarizes how active learning can integrate with batched perturbation experiments to achieve an efficient protocol.

3.3 Active Learning Algorithm

In this section, we propose an active learning algorithm (Algorithm 5) that integrates experimental design, deep learning, and biological analysis.

Let $\mathbf{1}_{\mathcal{I}} : [p] \to \{0, 1\}$ be an indicator function such that $\mathbf{1}_{\mathcal{I}}(i) = 1$ if and only if $i \in \mathcal{I}$. We use $\lfloor \cdot \rfloor$ to denote the floor function. For $t \in [T]$, let $\mathcal{H}(t)$ denote the history of experimental data observed before batch t. To be specific, we define $\mathcal{D}(t) \triangleq \left(\mathcal{N}(t), \mathcal{R}(t), \{A^n(t), \tilde{Y}^n(t)\}_{n \in \mathcal{N}(t) \cup \mathcal{R}(t)}\right)$ as the data observed in batch t. We then have $\mathcal{H}(t) \triangleq \{\mathcal{D}(\tau)\}_{\tau < t}$. Let $w_i(t)$ be the number of cells that have received lentiviruses carrying gene i before batch t. We call $w(t)$ an *observation frequency vector*. Let $1 \leq M_{\text{init}} \ll K$ be the minimum number of cells that we want to measure for each lentivirus type in the first batch of experiments.

Algorithm 5 Combinatorial Pure Exploration in Batched Experiments

Input: $K \in \mathbb{Z}_+$, $M_{\text{init}} \in \mathbb{Z}_+$, $T \in \mathbb{Z}_+$
Output: A
$\mathcal{H}(1) \leftarrow \emptyset$. ▷ Initialization
$w_i(1) \leftarrow 0$ for $i \in [p]$.
$\lambda_i(1) \leftarrow \mathbf{1}_{\mathcal{I}}(i) \ln(K/(K - M_{\text{init}}))$ for $i \in [p]$.
for $t = 1, \ldots, T$ **do**
    Experiment with $\lambda(t)$ and record data ▷ Perturbation experiment
    $\mathcal{D}(t) \leftarrow \left(\lambda(t), \mathcal{N}(t), \mathcal{R}(t), \{A^n(t), \tilde{Y}^n(t)\}_{n \in \mathcal{N}(t) \cup \mathcal{R}(t)}\right)$.
    $\mathcal{H}(t+1) \leftarrow \mathcal{H}(t) \cup \mathcal{D}(t)$.
    $w_i(t+1) \leftarrow w_i(t) + \sum_{n \in \mathcal{N}(t) \cup \mathcal{R}(t)} A_i^n(t)$ for $i \in \mathcal{I}$.
    $\rho(t+1) \leftarrow \texttt{Analyzer}\left(\mathcal{H}(t+1)\right)$. ▷ Frequency analysis
    $\hat{d}(t+1) \leftarrow \texttt{Model}\left(\mathcal{H}(t+1)\right)$. ▷ Distance estimation
    $J(t+1) \leftarrow \texttt{Oracle}\left(\rho(t+1), \hat{d}(t+1), w(t+1)\right)$. ▷ TF scoring
    $\lambda(t+1) \leftarrow \texttt{Solver}\left(J(t+1)\right)$. ▷ MOI optimization
end for
$A \leftarrow \operatorname{argmin}_{a \in \mathcal{F}} \rho(T+1)(a) \hat{d}(T+1)(a)$. ▷ Perturbation recommendation

3.3.1 Initialization

We use the first batch of experiments as an initialization step where we add all lentiviruses at the same MOI, which implies that each type of lentiviruses infects the same number of cells in expectation. Since we hope to estimate the perturbation effects of all lentiviruses to similar accuracy during initialization, we perturb K cells in the first batch of experiments and sequence all these cells without the enrichment step. Other batches of experiments are conducted as described in steps ii–v in Section 3.2.2 with CD34 enrichment and differential sampling. In order to measure at least M_{init} cells in expectation for each type of lentiviruses during initialization where $M_{\text{init}} \ll K$, we need $\left(1 - e^{-\lambda_i(1)}\right) K \geq M_{\text{init}}$ for $i \in \mathcal{I}$, which simplifies to $\lambda_i(1) \geq \ln(K/(K - M_{\text{init}}))$. We can show that the probability of a cell receiving more than one virion increases with the MOI of each lentivirus. Hence, minimizing the average fraction of multi-perturbed cells subject to the constraint of getting at least M_{init} cells in expectation for each lentivirus type gives $\lambda_i(1) = \mathbf{1}_{\mathcal{I}}(i) \ln(K/(K - M_{\text{init}}))$ for $i \in [p]$. Let $\Lambda > 0$ be an upper bound on the total MOI $\sum_{i \in \mathcal{I}} \lambda_i$ to prevent cells from generally dying due to excessive perturbation. Solving $\sum_{i \in \mathcal{I}} \lambda_i \leq \Lambda$, we get $M_{\text{init}} \leq K\left(1 - e^{-\Lambda/|\mathcal{I}|}\right)$.

3.3.2 Frequency Analysis

Let `Analyzer` be a subroutine that infers the desirability of a perturbation from its frequency in experimental data. To be specific, let $\rho : \{0,1\}^p \to [0,1]$ be a penalty function constructed by `Analyzer`. We assume that, for any perturbation a, we have $\rho(a) \approx 0$ if $\mathbb{E}_{\mathcal{P}^a}[b(f(X,a))] \approx 1$ and $\rho(a) \approx 1$ if $c(a) \approx 1$.[1] Intuitively, the penalty of a is high if a often kills cells and the penalty is low if a tends to induce the partial success of reprogramming.

We estimate cell death using the difference between predicted and actual cell counts. For example, the lethality penalty for the combination of TFs that overexpress genes i and $i' \neq i$ is given by $\mathbb{P}(A_i = 1)\,\mathbb{P}(A_{i'} = 1) - \mathbb{P}(A_i = 1, A_{i'} = 1)$. Under

[1] The conditions $\mathbb{E}_{\mathcal{P}^a}[b(f(X,a))] \approx 1$ and $c(a) \approx 1$ cannot be satisfied simultaneously by any perturbation because cell death precedes transcription as described in Section 3.2.2.

the assumption that the Poisson distributions of different lentiviruses are mutually independent, the lethality penalty equals 0 if the combination is nontoxic and positive otherwise. In principle, the penalty is an increasing function of the lethality of a perturbation. We estimate the penalty considering one batch of experiments at a time and average over all experiments with weights being proportional to the total cell counts of each experiment.

One method of incorporating enrichment information into the penalty is subtracting the relative frequency of each perturbation within the enriched set from its lethality penalty. Intuitively, the more often we observe a perturbation in $CD34^+$ cells, the more likely it can reprogram cells to the target cell type, and hence it should have a lower penalty. With the aforementioned example, the penalty now becomes $\mathbb{P}(A_i = 1)\,\mathbb{P}(A_{i'} = 1) - \mathbb{P}(A_i = 1, A_{i'} = 1) - P(A_i = 1, A_{i'} = 1 \mid b(f(X, A)) = 1)$. After computing the weighted average penalty over all experiments, we normalize the values to $[0, 1]$.

3.3.3 Distance Estimation

We use `Model` to represent a deep learning model that maps experimental data to an estimator $\hat{d}$ of d. In other words, we have $\hat{d}(a) \approx d(\mathcal{P}^a, \mathcal{Q})$ for any S-sparse perturbation a. There are numerous machine learning methods that can serve this purpose, including the neural tangent kernel (NTK) [112] and the compositional perturbation autoencoder (CPA) [90]. We use CPA for distance estimation in this chapter as it learns an interpretable linear model in the latent space where the effect of overexpressing a combination of genes is equal to the sum of the effects of perturbing each gene separately. The approach represents each TF as an embedding in the latent space, which allows us to predict cell behavior subject to novel combinations. We measure cosine similarity and hence the distance function d using perturbation embeddings. Fig. 3-4 provides an example of distance estimation in the CPA latent space. We show numerically in Section 3.4 that CPA outperforms benchmarks and provides TF embeddings that are consistent with biology. We defer the model assumptions, the mathematical formulation, and implementation details of CPA to Section 3.6.

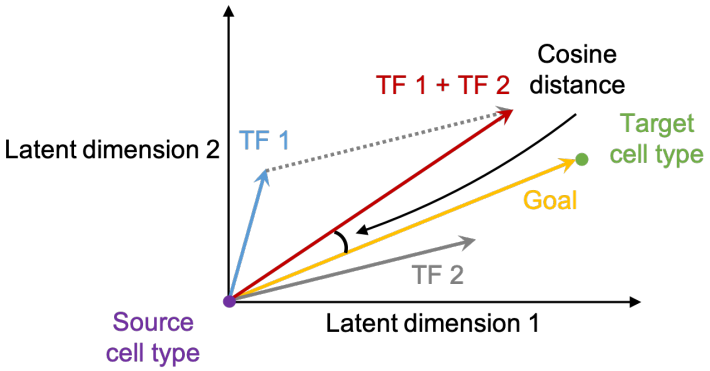

Figure 3-4: Example of distance estimation in the two-dimensional latent space of CPA for a pair of TFs.

3.3.4 TF Scoring

Having learned the penalty function ρ and the cost function $\hat{d}$, we call a process named `Oracle` to rate TF genes, which will determine the MOI vector for the next batch of experiments. To be precise, `Oracle` evaluates the cost function $J : \mathcal{I} \to \mathbb{R}$ given by

$$J(i) \triangleq \min_{a \in \mathcal{H}:\, a_i = 1} \rho(a)\hat{d}(a) - u(a^*(i), w), \tag{3.4}$$

where $a^*(i) \triangleq \operatorname{argmin}_{a \in \mathcal{H}:\, a_i = 1} \rho(a)\hat{d}(a)$. With slight abuse of notation, we denote the output of `Oracle` by J in Algorithm 5. For computational efficiency, we restrict the evaluation of ρ and $\hat{d}$ to the history of experimental data where $a \in \mathcal{H}$ means that perturbation a has been observed in past experiments. Given an observation frequency vector w, we introduce a term $u(a, w) > 0$ for each perturbation a that measures estimation uncertainty which decreases with the number of times that the genes overexpressed by a have been sequenced. Intuitively, we should overexpress a TF gene in the next batch of experiments if at least one of the following conditions is satisfied: (i) there is insufficient data about this TF; (ii) perturbations that include this TF rarely cause cell death and can drive the cell identity close to the target cell type. These two conditions embody the importance of exploration and exploitation, respectively. Inspired by the multi-armed bandit model, a simple yet powerful framework for sequential decision making under uncertainty [11, 85, 86, 114], we propose to balance the exploration–exploitation trade-off using Equation (3.4).

In the literature of combinatorial bandits [35, 36, 57, 83, 84, 96, 142, 144, 148, 152], the uncertainty term is defined for combinatorial actions as well as single actions. For example, $u_{\text{reg}}(a, w) = \beta\sqrt{\ln(tK)/a^\top w}$ with some constant coefficient $\beta > 0$ resembles the confidence radius of the combinatorial upper confidence bound (CUCB) algorithm [36]. CUCB is optimal in terms of cumulative regret for combinatorial multi-armed bandit problems with general reward functions [36, 84]. For the non-combinatorial pure exploration problem (equivalently, $S = 1$), the upper confidence bound exploration (UCB-E) algorithm using an uncertainty term similar to $u_{\text{exp}}(a, w) = \beta\sqrt{(TK - |\mathcal{I}|)/a^\top w}$ achieves optimal probability of error, namely, the probability that the recommendation at the end of exploration is suboptimal [10]. Although the optimality of UCB-E requires β to depend on the unknown hardness of the problem, tuning β empirically works well in practice [10]. For the combinatorial pure exploration problem (equivalently, $S \in \mathbb{Z}_+$), the combinatorial lower-upper confidence bound (CLUCB) algorithm also uses a confidence radius similar to $u_{\text{exp}}(a, w)$ and achieves optimality for many action sets [35]. Similar to UCB-E, CLUCB needs an unobserved hardness measure of the problem, which can be resolved by tuning β empirically. The uncertainty term u controls the level of exploration of the active learning algorithm. The pure exploration problem needs more exploration than the setting of cumulative regret minimization [10, 29]. In the limit of $u \to \infty$, the algorithm reduces to uniform sampling which is suboptimal for the regime studied in this chapter where the action set is large yet the time horizon is short [10, 29]. Although the conditions under which multi-armed bandit algorithms are studied, such as the monotonicity of the reward function, are unrealistic for biological applications, we demonstrate how the core idea of balancing exploration and exploitation can help design active learning algorithms for efficient experimentation.

The active learning problem considered in this chapter can be described as combinatorial fixed-budget pure exploration of batched bandits with incomplete information, which has not been investigated in the literature. However, parts of this topic have been well studied separately. We hope that our work underlines the importance of addressing the aforementioned challenges simultaneously. Since our setting is a

pure exploration problem, we use an uncertainty term similar to $u_{\text{exp}}(a, w)$, defined as $u(a, w) = \beta\sqrt{(TK - |\mathcal{I}|)/(a^\top w + 1)}$ for measuring estimation uncertainty. The only difference is that we add 1 to the frequency sum $a^\top w$ so that $u(a, w) < \infty$. In contrast to the ideal bandit setting where one always observes the rewards of the actions taken, a perturbation can be missing in experimental data due to cell death and random sampling. For this reason, we modify the uncertainty term to make it bounded.

3.3.5 MOI Optimization

Given a constant $\alpha \in \mathbb{R}$, we can find an optimal MOI vector by solving the following optimization problem:

$$\begin{aligned} \min_{\lambda} \quad & \sum_{i \in \mathcal{I}} (\lambda_i + J(i) - \alpha)^2 \\ \text{s.t.} \quad & \sum_{i \in \mathcal{I}} \lambda_i \;\; \leq \;\; \Lambda \\ & \lambda_i \;\; \geq \;\; 0 \qquad\qquad \text{if } i \in \mathcal{I} \\ & \lambda_i \;\; = \;\; 0 \qquad\qquad \text{otherwise.} \end{aligned} \tag{3.5}$$

At a high level, the better a TF gene is according to Equation (3.4), the higher we want to set the MOI of the corresponding lentivirus. We define $\gamma \in \mathbb{R}^p$ as $\gamma_i = \mathbf{1}_{\mathcal{I}}(i)(\alpha - J(i))$ for $i \in [p]$. We can consider the optimization problem (3.5) as projecting γ on a set of MOI vectors that satisfy the upper bound Λ. Since the feasible set is nonempty, closed, and convex, and the Euclidean norm is strictly convex, we conclude that, for any $\{J(i)\}_{i \in \mathcal{I}}$ and α, (3.5) has a unique optimal solution [27]. We can also show that the projection is a continuous function [27]. Lemma 7 provides the closed-form solution to (3.5), which can be constructed efficiently as discussed in the proof. Fig. 3-5 illustrates Lemma 7 with randomly generated Gaussian costs.

Lemma 7. *If* $\sum_{i \in \mathcal{I}} \max(\alpha - J(i), 0) \leq \Lambda$, *then* $\lambda_i^* = \mathbf{1}_{\mathcal{I}}(i) \max(\alpha - J(i), 0)$ *for* $i \in [p]$. *Otherwise, there exists a nonempty subset* $\mathcal{I}_0 \subseteq \mathcal{I}$ *such that the solution*

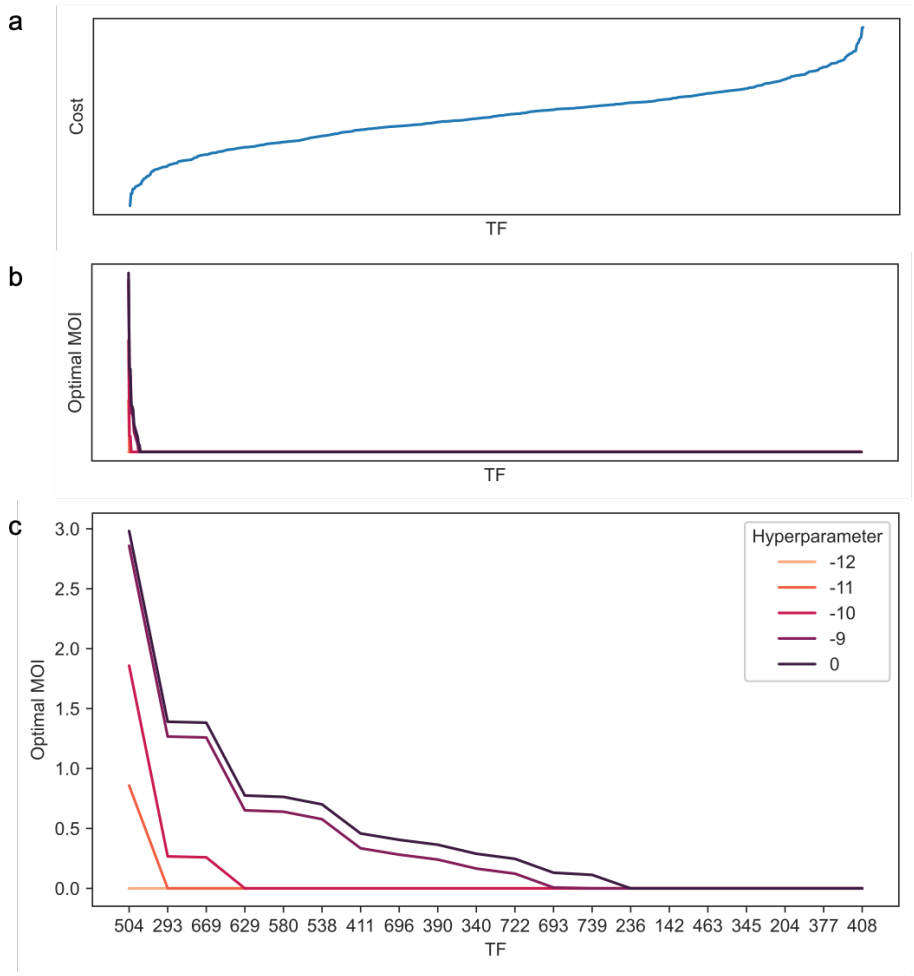

Figure 3-5: Example of the MOI optimization problem (3.5) with Gaussian cost. a, Costs of 800 TFs are independently sampled from the Gaussian distribution with a mean of 0 and a standard deviation of 4. b, Optimal MOI vectors are solved for different values of the hyperparameter α. c, The optimal MOI vector increases with α. For a fixed α, the lower the cost of a TF, the higher the optimal MOI.

to (3.5) *is given by, for* $i \in [p]$,

$$\lambda_i^* = \mathbf{1}_{\mathcal{I}_0}(i) \left[\frac{\Lambda + \sum_{i' \in \mathcal{I}_0} J(i')}{|\mathcal{I}_0|} - J(i) \right]. \tag{3.6}$$

Proof. If $\sum_{i \in \mathcal{I}} \max(\alpha - J(i), 0) \leq \Lambda$, then $\max(\gamma, 0)$ is feasible and hence $\lambda^* = \max(\gamma, 0)$.[2] For the rest of the proof, we assume that $\max(\gamma, 0)$ is infeasible and

[2]Maximization is applied element-wise.

solve (3.5) using the Karush–Kuhn–Tucker (KKT) conditions. For simplicity of notation, we drop the equality constraint in (3.5) and optimize over the $|\mathcal{I}|$-dimensional subspace of $[0, \infty)^p$ that corresponds to TF genes. The simplified problem with $\lambda_i = 0$ for $i \notin \mathcal{I}$ is equivalent to (3.5) and hence a convex optimization problem. The KKT conditions are necessary and sufficient for the optimality of convex optimization problems [27]. Thus, we use KKT conditions to find the unique globally optimal solution to (3.5).

Let $\nu \in \mathbb{R}^{|\mathcal{I}|+1}$ be the Lagrange multiplier vector associated with the inequality constraints. By KKT conditions, λ^* and ν^* are primal and dual optimal if and only if the following equations are satisfied simultaneously:

$$\sum_{i \in \mathcal{I}} \lambda_i^* \leq \Lambda, \tag{3.7}$$

$$\lambda_i^* \geq 0, \quad i \in \mathcal{I}, \tag{3.8}$$

$$\nu_i^* \geq 0, \quad i \in \{0\} \cup \mathcal{I}, \tag{3.9}$$

$$\nu_0^* \left[\left(\sum_{i \in \mathcal{I}} \lambda_i^* \right) - \Lambda \right] = 0, \tag{3.10}$$

$$\nu_i^* \lambda_i^* = 0, \quad i \in \mathcal{I}, \tag{3.11}$$

$$2(\lambda_i^* + J(i) - \alpha) + \nu_0^* - \nu_i^* = 0, \quad i \in \mathcal{I}. \tag{3.12}$$

Equations (3.7) and (3.8) represent primal feasibility. Equation (3.9) is dual feasibility. Equations (3.10) and (3.11) correspond to complementary slackness. Equation (3.12) indicates stationarity of the Lagrangian. Equation (3.12) gives

$$\lambda_i^* = \frac{\nu_i^* - \nu_0^*}{2} - J(i) + \alpha, \quad i \in \mathcal{I}. \tag{3.13}$$

Using Equations (3.11) and (3.13), we get

$$\nu_i^* \left(\frac{\nu_i^* - \nu_0^*}{2} - J(i) + \alpha \right) = 0, \quad i \in \mathcal{I}. \tag{3.14}$$

We define $\mathcal{I}_0 \subseteq \mathcal{I}$ as $\mathcal{I}_0 \triangleq \{\ i \in \mathcal{I} \mid \nu_i^* = 0\ \}$.

We first assume that $\nu_0^* = 0$ and derive a contradiction. Equation (3.14) implies that $\nu_i^* = 0$ or $\nu_i^* = 2(J(i) - \alpha)$ for all $i \in \mathcal{I}$. We fix any $i \in \mathcal{I}$. If $J(i) \leq \alpha$, then dual feasibility (Equation (3.9)) indicates that $\nu_i^* = 0$. If $J(i) > \alpha$, then primal feasibility (Equation (3.8)) together with Equation (3.13) requires that $\nu_i^* = 2(J(i) - \alpha)$. Hence, we have $\lambda^* = \max(\gamma, 0)$, which contradicts our assumption that $\max(\gamma, 0)$ is infeasible.

We now assume that $\nu_0^* \neq 0$. If $\mathcal{I}_0 = \emptyset$, then Equation (3.11) gives $\lambda_i^* = 0$ for all $i \in \mathcal{I}$. It follows that $\nu_0^* \left[\left(\sum_{i\in\mathcal{I}} \lambda_i^*\right) - \Lambda\right] = -\nu_0^*\Lambda \neq 0$, which contradicts Equation (3.10). Thus, $\mathcal{I}_0 \neq \emptyset$. We define $\chi \triangleq \left(\Lambda + \sum_{i\in\mathcal{I}_0} J(i)\right) / |\mathcal{I}_0|$. By Equations (3.10), (3.13), and (3.14), we get $\nu_0^* = 2(\alpha - \chi)$. Together with Equation (3.14), we get $\nu_i^* = \mathbf{1}_{\mathcal{I}\setminus\mathcal{I}_0}(i)(2J(i) - 2\chi)$ for $i \in \mathcal{I}$. Using Equation (3.13), we have $\lambda_i^* = \mathbf{1}_{\mathcal{I}_0}(i)(\chi - J(i))$ for $i \in \mathcal{I}$. Since $\nu_0^* \neq 0$, Equation (3.10) implies that Equation (3.7) is satisfied. Primal feasibility (Equation (3.8)) requires $\chi \geq \max_{i\in\mathcal{I}_0} J(i)$. Dual feasibility (Equation (3.9)) needs $\chi \leq \alpha$ and $\chi \leq \min_{i\in\mathcal{I}\setminus\mathcal{I}_0} J(i)$. Since (3.5) has a unique optimal solution, there exists a unique subset $\mathcal{I}_0 \subseteq \mathcal{I}$ such that

$$\max_{i\in\mathcal{I}_0} J(i) \leq \chi \leq \min\left(\alpha, \min_{i\in\mathcal{I}\setminus\mathcal{I}_0} J(i)\right). \tag{3.15}$$

The threshold conditions of Equation (3.15) allow us to search for $\mathcal{I}_0$ efficiently. Let $\pi \in \mathbb{Z}_+^{|\mathcal{I}|}$ be a permutation of $\mathcal{I}$ such that $J(\pi_1) \leq J(\pi_2) \leq \cdots \leq J(\pi_{|\mathcal{I}|})$. A necessary condition for Equation (3.15) is

$$\max_{i\in\mathcal{I}_0} J(i) \leq \min_{i\in\mathcal{I}\setminus\mathcal{I}_0} J(i). \tag{3.16}$$

Thus, we can initialize $\hat{\mathcal{I}}_0 = \{\pi_1\}$ and include $\pi_2, \pi_3, \pi_4, \ldots$ one at a time until Equation (3.15) is satisfied. □

3.3.6 Perturbation Recommendation

After T batches of experiments, Algorithm 5 recommends the best perturbation based on the data collected. For this purpose, we minimize the exploitation term of the

`Oracle` cost function (Equation (3.4)) using penalty and distance estimated on all experimental data. In other words, the recommended perturbation is given by

$$A \triangleq \text{argmin}_{a \in \mathcal{F}} \rho(T+1)(a)\hat{d}(T+1)(a). \tag{3.17}$$

Since recommendation is an exploitation step, there is no need to account for estimation uncertainty in Equation (3.17). We provide numerical experiments in Section 3.4 to validate recommendation results with gene expression data.

3.4 Experiments

In this section, we conduct numerical experiments using a private dataset collected by the Zhang laboratory of the Broad Institute of MIT and Harvard. The Perturb-seq technique [47, 71, 113] was used to gather single-cell RNA sequencing data together with perturbation information. The dataset includes ten experiments, totaling 155501 cells and 31915 genes. We use the subset of the data that does not correspond to colony-forming units, which comprises 132578 cells from nine experiments. Within this subset, one experiment includes 4776 target cells. Fig. 3-6 summarizes the other eight experiments that contain source cells and perturbed cells.

3.4.1 Distance Estimation

In this section, we present the results of applying CPA to the screen data. We show that CPA outperforms benchmarks in terms of prediction accuracy. In addition, we verify that CPA learns TF embeddings that are consistent with biology.

We remove batch effects by setting the sample label as the covariate. During data preprocessing, we keep only 4914 differentially expressed (DE) genes that best distinguish the target cell type from the source cell type. These DE genes include 13 target identity genes in addition to top DE genes that are common to the screen data and the human cell landscape dataset [64]. We also remove labels with insufficient data. Moreover, we subsample source and target cells to balance class sizes by label.

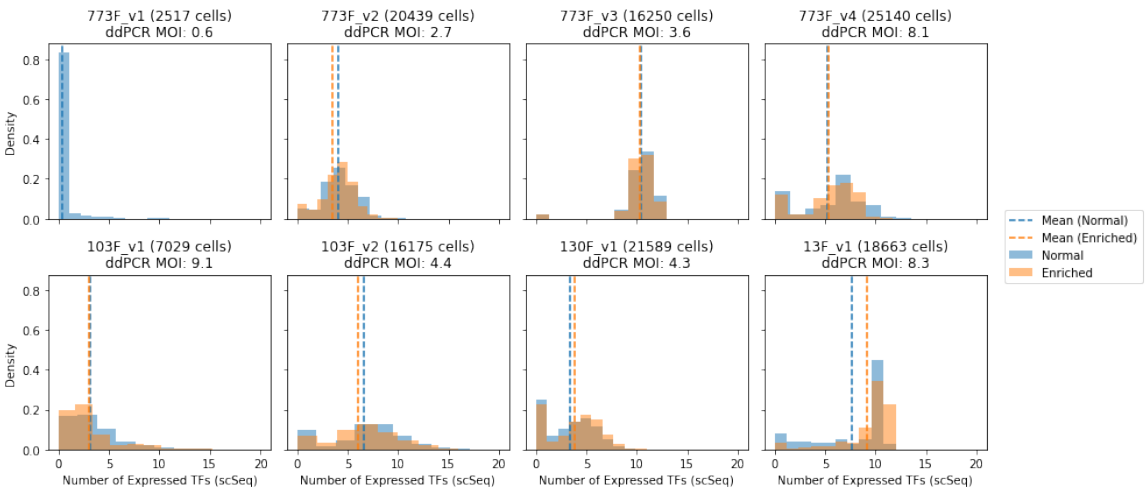

Figure 3-6: The distribution of the number of TFs received at the single-cell level by experiment. Except for `773F_v1`, each experiment has two samples, one corresponding to an unbiased sample of the cell population (called normal) and the other mainly consisting of enriched cells (referred to as enriched). The `773F_v1` experiment only has a normal sample.

We use several TF combinations as the out-of-distribution (OOD) set, which are excluded from the training set and the test set. The OOD set allows us to evaluate the extrapolation capability of CPA. After removing the OOD set, we randomly partition the rest of the data into a training set (80%) and a test set (20%). We use eleven train-test-OOD splits for cross-validation. The preprocessed data consists of 50220 cells with 410 unique perturbations.

We consider four benchmarks for evaluating the accuracy of mean prediction by CPA:

- `Mean`: Predicting mean gene expression of all cells.
- `Mean by TF`: Predicting mean gene expression of all TFs that comprise the given perturbation.
- `Mean by sample`: Predicting mean gene expression of the given sample.
- `Mean by sample and TF`: Predicting mean gene expression of all TFs that comprise the given perturbation, restricted to the given sample.

Fig. 3-7 shows that CPA outperforms all benchmarks on training, test, and OOD sets

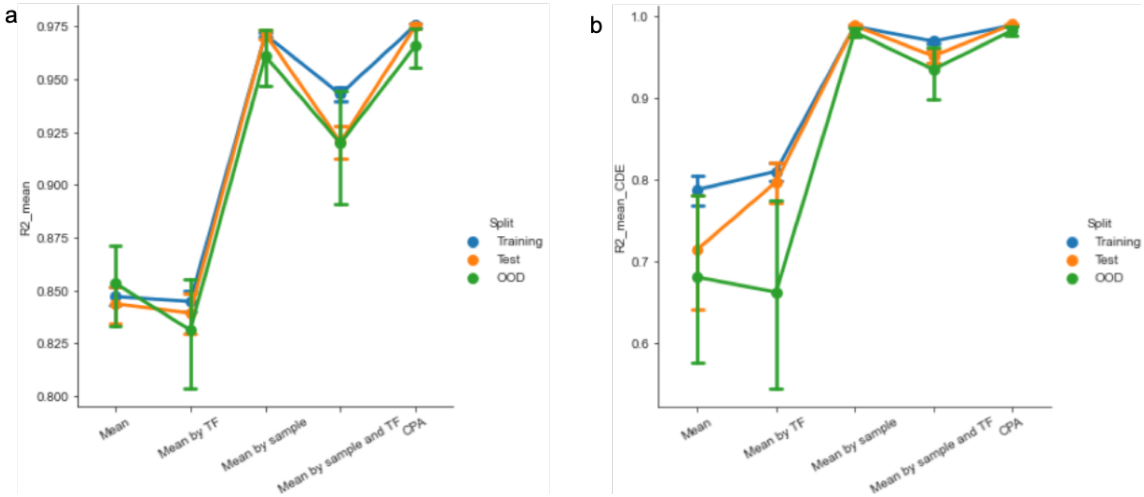

Figure 3-7: CPA outperforms all benchmarks on training, test, and OOD sets in terms of mean prediction accuracy. Bars correspond to a 95% confidence interval. a, The R^2 score on all genes. b, The R^2 score on top DE genes.

in terms of mean prediction accuracy. Table 3.1 lists the mean prediction accuracy of CPA and the benchmarks across splits. Although `mean by sample` runs a close second due to batch effects, this benchmark cannot serve the purpose of predicting perturbation effects. Fig. 3-7 implies that the advantage of CPA over `mean by sample and TF` which ranks third is statistically significant.

Next, we demonstrate that the TF embeddings learned by CPA are consistent with biology. We rank perturbations by their mean cosine similarity with the goal direction over all splits in the latent space, where the goal direction is the vector going from the source cell type to the target cell type. Due to the huge space of possible combinations, we only show top and bottom single TFs and pairs of TFs in Fig. 3-8. As shown in Fig. 3-8, qualitative results are consistent for top and bottom perturbations across different splits. In particular, the cosine similarity with the goal direction of top perturbations tends to exceed their mean cosine similarity with other perturbations, which indicates statistical significance. The TF that best aligns with the goal overexpresses TAL1, which is an identity gene for the target cell type. Similarly, the cosine similarity with the goal of bottom perturbations tends to be lower than their mean cosine similarity with other perturbations. We will give further evidence that the CPA embeddings agree with biology in Section 3.4.2.

Table 3.1: CPA outperforms all benchmarks on training, test, and OOD sets in terms of mean prediction accuracy.

Split	Model	All genes	Top DE genes
Training	Mean	0.847	0.788
	Mean by TF	0.845	0.810
	Mean by sample	0.971	0.988
	Mean by sample and TF	0.943	0.970
	CPA	**0.976**	**0.990**
Test	Mean	0.844	0.714
	Mean by TF	0.839	0.797
	Mean by sample	0.971	0.989
	Mean by sample and TF	0.920	0.952
	CPA	**0.975**	**0.990**
OOD	Mean	0.853	0.680
	Mean by TF	0.831	0.662
	Mean by sample	0.961	0.981
	Mean by sample and TF	0.920	0.935
	CPA	**0.966**	**0.983**

Figure 3-8: Ranking of single TFs and pairs of TFs by cosine similarity with the goal direction.

3.4.2 TF Scoring

In this section, we validate our proposed approach to scoring TFs. To be specific, we evaluate Equation (3.4) without the uncertainty term for each TF. This can be considered as running Algorithm 5 for one step. We hold out enriched samples and only use them for validation. Thus, the penalty only takes cell death into consideration. Fig. 3-9 implies that, among TFs that do not kill cells, the more likely a TF leads to enrichment, the lower the score. Similarly, there is an inverse relationship between (i) enrichment and distance, (ii) enrichment and penalty. In summary, our proposed approach can detect TFs that tend to lead to the partial success of reprogramming. The results can be further improved by replacing CPA with deep learning models that have increased prediction accuracy. The modularity of Algorithm 5 enables changing the deep learning component with ease.

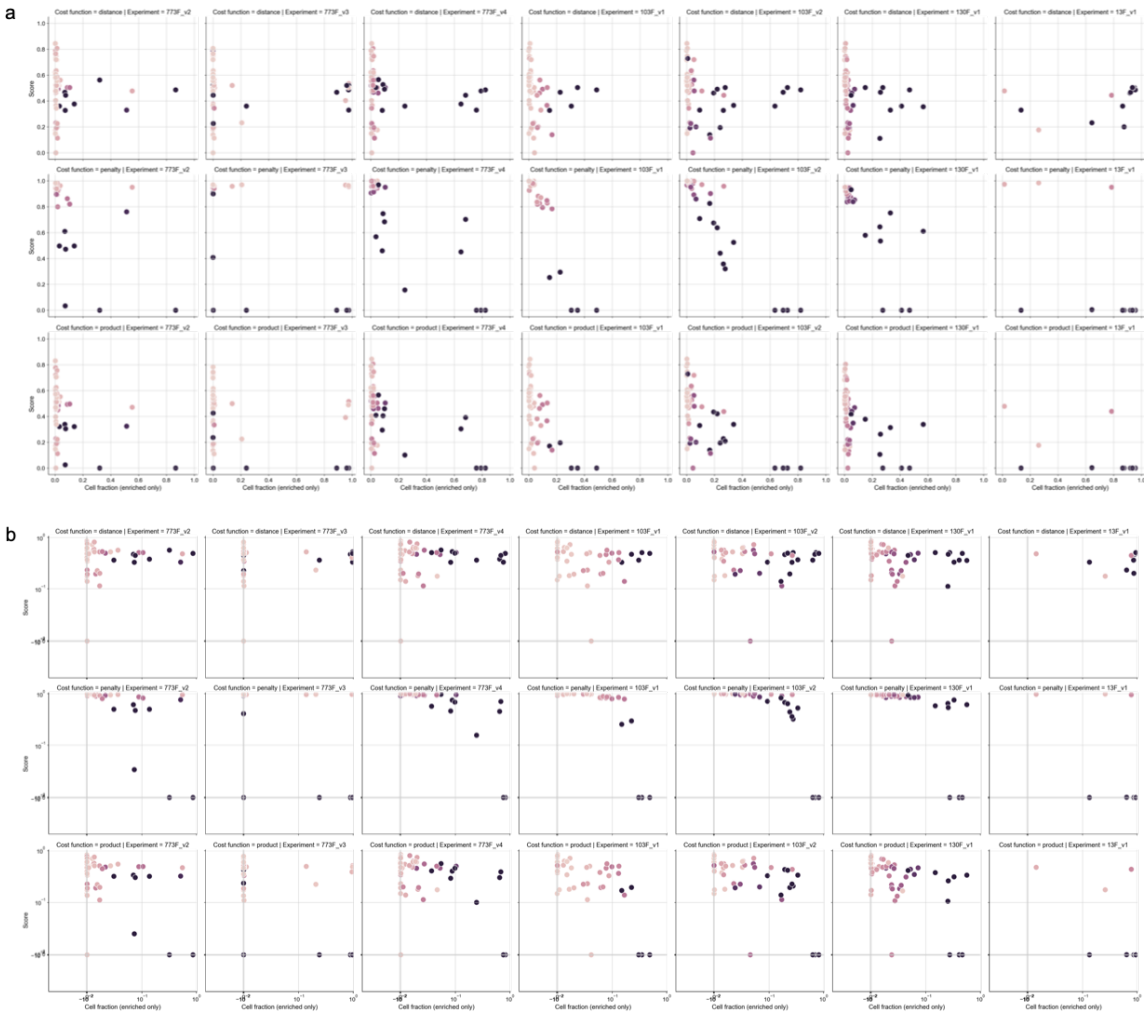

Figure 3-9: Controlling for cell survival, the more likely a TF leads to enrichment, the lower the score. Each point is a TF. The x-axis is the relative frequency in the enriched sample. From top to bottom, the y-axis are distance, penalty, and the TF score equal to the product, respectively. Color denotes cell survival, with higher cell counts in the normal sample being darker. Columns correspond to experiments. a, Both axes on a linear scale. b, Both axes on a symmetric logarithmic scale with a small linear interval around zero.

3.5 Discussion

In this chapter, we have formulated a framework to study the problem of optimizing perturbations for cell reprogramming in batched experiments. We have proposed an active learning algorithm that directly optimizes over combinations of TFs. Our method combines deep learning-based distance estimation and biology-based frequency analysis into one model. Although the complexity of the problem prohibits theoretical guarantees on the entire procedure, our approach is built upon multi-armed bandit algorithms which are optimal under certain conditions. We have demonstrated the success of our approach on data collected using large-scale perturbation screening that is pushing the frontiers of genetics research. Numerical experiments show that our method can identify TFs that are promising for successful reprogramming.

Our work has opened up several exciting directions for future research. First, it is valuable to demonstrate the ability of our approach to recommend good combinations of TFs beyond one-step experiments. This can be achieved through designing numerical experiments that simulate the algorithm for a few batches. Second, it is important to formulate the MOI optimization in a way that makes the sparsity of the solution depend explicitly on the number of batches. Third, benchmarking our algorithm against some commonly used heuristics can give insights into the value added by the combination-based active learning approach. In addition, it is important to develop deep learning models that are empirically and provably good. Finally, the ultimate goal is to successfully deploy the proposed approach in laboratories.

3.6 Supplementary Information

3.6.1 Compositional Perturbation Autoencoder

Model Assumptions

CPA needs a dataset of single-cell measurements. Let $a \in \{0,1\}^p$ be the perturbation to a cell. Let $y \in \mathbb{R}^p$ be the gene expression level of the cell subject to perturbation. Let $k \in \mathbb{Z}_+$ be the number of experiments conducted. We use a one-hot vector $v \in \{0,1\}^k$ to represent the experiment of the cell measurement. The covariate v allows CPA to control for batch effects due to variation in experimental procedures. Given a collection of tuples $\mathcal{C} \triangleq \{(y,a,v)\}$, CPA assumes the dataset to be produced by an unknown generative model described as

$$\begin{aligned} z &= z^{\text{basal}} + V^{\text{pert}} a + V^{\text{cov}} v, \\ y &\sim \mathbf{P}\left(\cdot \mid z\right). \end{aligned} \tag{3.18}$$

The key assumption of CPA is that there exists an r-dimensional latent space such that perturbation effects can be linearly disentangled from inherent stochasticity in gene expression and batch effects. Although this assumption remains to be validated, CPA outperforms benchmarks on our screen data as shown in Section 3.4.1 and hence the model is reasonable. The basal latent state $z^{\text{basal}} \in \mathbb{R}^r$ captures the inherent stochasticity in gene expression of the source cell. The matrices $V^{\text{pert}} \in \mathbb{R}^{r \times p}$ and $V^{\text{cov}} \in \mathbb{R}^{r \times k}$ describe perturbation response and batch effects, respectively. CPA assumes that z^{basal} is independent of a and v. Conditional on the latent state z, the gene expression level y of the perturbed cell is sampled from an unknown decoding distribution $\mathbf{P}\left(\cdot \mid z\right)$. The only observations of Equation (3.18) are (y,a,v).

The general setting of CPA allows continuous perturbations such as a combination of different drugs at various doses [90]. Moreover, it is straightforward to consider more than one covariate [90]. In this section, we have introduced a simplified version of CPA as described by Equation (3.18), which is sufficient for distance estimation.

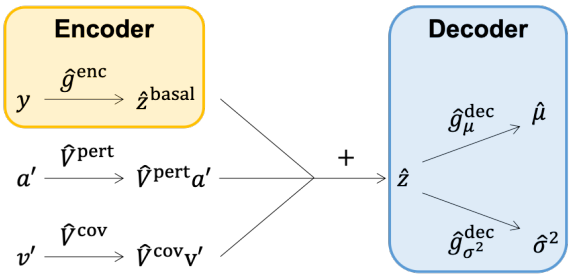

Figure 3-10: The prediction procedure of CPA.

Prediction

CPA learns the model (Equation (3.18)) to make counterfactual predictions, namely, what would the gene expression be if the cell were given a different perturbation under some other condition? In order to answer this question, CPA first encodes the observed gene expression y into a basal latent state $\hat{z}^{\text{basal}}$ and then estimates the perturbed latent state $\hat{z}$. Let $\mu(z)$ and $\sigma^2(z)$ be the mean and variance vectors of $\mathbf{P}(\cdot \mid z)$ where the ith entry corresponds to gene i in y. In other words, we have $\mu(z) = (\mu_1(z), \ldots, \mu_p(z))$ where $\mu_i(z) \triangleq \mathbb{E}[Y_i \mid z]$. Similarly, $\sigma^2(z) = \left(\sigma_1^2(z), \ldots, \sigma_p^2(z)\right)$ where $\sigma_i^2(z) \triangleq \text{Var}(Y_i \mid z)$. Finally, CPA decodes $\hat{z}$ to estimate the mean and variance vectors $\hat{\mu}$ and $\hat{\sigma}^2$ for the queried perturbation a' and covariate v'. Let $\hat{g}^{\text{enc}}$ be the encoder. We use $\hat{g}_{\mu}^{\text{dec}}$ and $\hat{g}_{\sigma^2}^{\text{dec}}$ to denote the decoders for mean and variance, respectively. Let $\hat{V}^{\text{pert}}$ be the estimate for V^{pert}. We define $\hat{V}^{\text{cov}}$ similarly. The prediction procedure can be summarized as follows and illustrated in Fig. 3-10:

$$\begin{aligned}
\hat{z}^{\text{basal}} &= \hat{g}^{\text{enc}}(y), \\
\hat{z} &= \hat{z}^{\text{basal}} + \hat{V}^{\text{pert}}a' + \hat{V}^{\text{cov}}v', \\
\hat{\mu} &= \hat{g}_{\mu}^{\text{dec}}(\hat{z}), \\
\hat{\sigma}^2 &= \hat{g}_{\sigma^2}^{\text{dec}}(\hat{z}).
\end{aligned} \tag{3.19}$$

We will discuss how the estimates in Equation (3.19) are optimized during training in the next section.

Training

In order to accurately reproduce the observed gene expression, CPA minimizes a reconstruction loss defined as the Gaussian negative log-likelihood given by

$$L^{\text{recon}}\left(\hat{V}^{\text{pert}}, \hat{V}^{\text{cov}}, \hat{g}^{\text{enc}}, \hat{g}_{\mu}^{\text{dec}}, \hat{g}_{\sigma^2}^{\text{dec}}\right) \triangleq \sum_{(y,a,v)\in\mathcal{C}} \frac{1}{p} \sum_{i=1}^{p} \left[\frac{1}{2} \ln \zeta\left(\hat{\sigma}_i^2\right) + \frac{\left(y_i - \hat{\mu}_i\right)^2}{2\zeta\left(\hat{\sigma}_i^2\right)}\right],$$

where $\zeta\left(\hat{\sigma}_i^2\right) \triangleq \ln\left(\exp\left(\hat{\sigma}_i^2 + 10^{-3}\right) + 1\right) \approx \hat{\sigma}_i^2$ guarantees positivity of the variance and improves numerical stability.

The key assumption of CPA to extrapolate unseen combinatorial responses is that perturbation effects can be separated from inherent stochasticity in gene expression and batch effects. To this end, two adversaries are introduced, referred to as $\hat{h}^{\text{pert}} : \mathbb{R}^r \to \mathbb{R}^{|\mathcal{I}|}$ and $\hat{h}^{\text{cov}} : \mathbb{R}^r \to \mathbb{R}^k$. Specifically, one adversary predicts the likelihood of perturbation $\hat{q}^{\text{pert}} \triangleq \hat{h}^{\text{pert}}\left(\hat{z}^{\text{basal}}\right)$ from the estimated basal latent state, where $\hat{q}_i^{\text{pert}}$ is the estimated probability that gene i is overexpressed in the cell queried. The other adversary predicts the covariate likelihood $\hat{q}^{\text{cov}} \triangleq \hat{h}^{\text{cov}}\left(\hat{z}^{\text{basal}}\right)$ with $\hat{q}^{\text{cov}}$ being a probability vector over all experiments. The goal of these adversaries is to minimize the cross-entropy loss given by

$$L^{\text{adv}}\left(\hat{h}^{\text{pert}}, \hat{h}^{\text{cov}}, \hat{g}^{\text{enc}}\right) \triangleq \frac{1}{|\mathcal{C}|} \sum_{(y,a,v)\in\mathcal{C}} \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \left[-a_i \ln \hat{q}_i^{\text{pert}} - (1 - a_i) \ln\left(1 - \hat{q}_i^{\text{pert}}\right)\right] + \frac{1}{|\mathcal{C}|} \sum_{(y,a,v)\in\mathcal{C}} - \ln \hat{q}^{\text{cov}}_{\text{argmax}_j v_j}.$$

In contrast to the adversaries, the autoencoder wants to disentangle perturbation effects and hence maximize the adversarial loss. The rivalry between the autoencoder and the adversaries can be considered as $\max_{\hat{g}^{\text{enc}}} \min_{\hat{h}^{\text{pert}}, \hat{h}^{\text{cov}}} L^{\text{adv}}$.

The autoencoder achieves gene expression reconstruction and perturbation effect disentanglement simultaneously by minimizing a weighted average of both losses, namely,

$$L \triangleq L^{\text{recon}} - \kappa L^{\text{adv}}, \tag{3.20}$$

where $\kappa > 0$ is a parameter that balances the adversarial loss and the reconstruction loss. CPA is trained by iterating between optimizing the autoencoder and optimizing the adversaries:

i (Autoencoder optimization) With $\left(\hat{h}^{\text{pert}}, \hat{h}^{\text{cov}}\right)$ fixed, sample $\mathcal{C}$ and find

$$\hat{V}^{\text{pert}}, \hat{V}^{\text{cov}}, \hat{g}^{\text{enc}}, \hat{g}_{\mu}^{\text{dec}}, \hat{g}_{\sigma^2}^{\text{dec}} = \operatorname{argmin} \left(L^{\text{recon}} - \kappa L^{\text{adv}}\right).$$

ii (Adversary optimization) With $\left(\hat{V}^{\text{pert}}, \hat{V}^{\text{cov}}, \hat{g}^{\text{enc}}, \hat{g}_{\mu}^{\text{dec}}, \hat{g}_{\sigma^2}^{\text{dec}}\right)$ fixed, sample $\mathcal{C}$ and find

$$\hat{h}^{\text{pert}}, \hat{h}^{\text{cov}} = \operatorname{argmin} L^{\text{adv}}.$$

Evaluation

In this section, we first present the mathematical derivations for the statistics used in CPA evaluation. We then provide pseudocode for computing the R^2 scores of mean and variance prediction in Algorithm 6.

We start with explaining the method of evaluating mean prediction. Conditional on a pair of perturbation and covariate (a, v), the distribution of the latent state z is that of the basal latent state z^{basal} shifted by a constant vector which depends on (a, v). Let $z^{\text{basal},1}, \ldots, z^{\text{basal},N}$ be independent and identically distributed samples of z^{basal}. Let $\overline{\mu}(a, v) \triangleq N^{-1} \sum_{n=1}^{N} \mu(z^n)$ where $z^n \triangleq z^{\text{basal},n} + V^{\text{pert}} a + V^{\text{cov}} v$. It then follows that

$$\begin{aligned}
\mathbb{E}\left[y \mid a, v\right] &= \mathbb{E}\left[\mathbb{E}\left[y \mid z\right] \mid a, v\right] \\
&= \mathbb{E}\left[\mu(z) \mid a, v\right] \\
&= \int \mu(z) \mathbf{P}\left(z^{\text{basal}}\right) dz^{\text{basal}} \\
&\approx \overline{\mu}(a, v).
\end{aligned}$$

Thus,

$$\overline{\hat{\mu}}(a, v) \approx \overline{\mu}(a, v) \approx \mathbb{E}\left[y \mid a, v\right] = \mathbb{E}\left[\overline{y}(a, v) \mid a, v\right]. \tag{3.21}$$

We now derive the formula for the assessment of variance prediction. Let $\overline{\sigma^2}(a,v) \triangleq N^{-1}\sum_{n=1}^{N}\sigma^2(z^n)$. We define a vector $\overline{\omega}(a,v) \in \mathbb{R}^p$ where

$$\overline{\omega}_i(a,v) \triangleq \frac{1}{N}\sum_{n=1}^{N}\left(\hat{\mu}_i(z^n) - \overline{\hat{\mu}}_i(a,v)\right)^2,$$

for $i \in [p]$. Thus,

$$\begin{aligned}
\mathrm{Var}\left(\mu_i(z) \mid a,v\right) &= \mathbb{E}\left[\left(\mu_i(z) - \mathbb{E}\left[\mu_i(z) \mid a,v\right]\right)^2 \,\middle|\, a,v\right] \\
&= \int \left(\mu_i(z) - \mathbb{E}\left[\mu_i(z) \mid a,v\right]\right)^2 \mathbf{P}\left(z^{\text{basal}}\right) dz^{\text{basal}} \\
&\approx \frac{1}{N}\sum_{n=1}^{N}\left(\mu_i(z^n) - \overline{\mu}_i(a,v)\right)^2 \\
&\approx \frac{1}{N}\sum_{n=1}^{N}\left(\hat{\mu}_i(z^n) - \overline{\hat{\mu}}_i(a,v)\right)^2 \\
&= \overline{\omega}_i(a,v).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathrm{Var}\left(y_i \mid a,v\right) &= \mathbb{E}\left[\mathrm{Var}\left(y_i \mid z\right) \mid a,v\right] + \mathrm{Var}\left(\mathbb{E}\left[y_i \mid z\right] \mid a,v\right) \\
&= \mathbb{E}\left[\sigma_i^2(z) \,\middle|\, a,v\right] + \mathrm{Var}\left(\mu_i(z) \mid a,v\right),
\end{aligned}$$

and

$$\mathbb{E}\left[\sigma_i^2(z) \,\middle|\, a,v\right] = \int \sigma_i^2(z)\mathbf{P}\left(z^{\text{basal}}\right) dz^{\text{basal}} \approx \overline{\sigma_i^2}(a,v).$$

We get that

$$\mathrm{Var}\left(y_i \mid a,v\right) \approx \overline{\sigma_i^2}(a,v) + \overline{\omega}_i(a,v).$$

Therefore,

$$\overline{\hat{\sigma}_i^2}(a,v) \approx \overline{\sigma_i^2}(a,v) \approx \mathrm{Var}\left(y_i \mid a,v\right) - \overline{\omega}_i(a,v) \approx \mathbb{E}\left[s_i^2(a,v) \,\middle|\, a,v\right] - \overline{\omega}_i(a,v). \quad (3.22)$$

Finally, we present pseudocode for computing the R^2 scores of mean and variance prediction in Algorithm 6, which is used for benchmarking in Section 3.4.1. At a high

level, CPA first computes R^2 scores of mean and variance prediction controlling for covariate and perturbation, and then averages each score over all groups. We recall that the autoencoder is defined by $\left(\hat{V}^{\text{pert}}, \hat{V}^{\text{cov}}, \hat{g}^{\text{enc}}, \hat{g}_{\mu}^{\text{dec}}, \hat{g}_{\sigma^2}^{\text{dec}}\right)$. For simplicity of notation, we use `autoencoder` to represent the model in Algorithm 6. By Equation (3.21), given any (a, v), there should be a good correlation between $\overline{\hat{\mu}}(a, v)$ and $\overline{y}(a, v)$ and hence lines 12 and 13 of Algorithm 6 are justified. Similarly, Equation (3.22) explains lines 14 and 15. We note that lines 14 and 15 correct the evaluation in [90] that mistakenly uses $s^2(a, v)$ and $\overline{\hat{\sigma}^2}(a, v)$ to calculate the R^2 scores for variance. The output of Algorithm 6 comprises the R^2 score of mean using all genes (R^2_{mean}), the R^2 score of mean using top DE genes ($R^2_{\text{mean,DE}}$), the R^2 score of variance using all genes (R^2_{var}), and the R^2 score of variance using top DE genes ($R^2_{\text{var,DE}}$).

Algorithm 6 R^2 Scores of Mean and Variance Prediction by CPA

1: **Input:** `autoencoder`, $\texttt{cells}_{\texttt{perturbed}}$, $\texttt{cells}_{\texttt{source}}$
2: **Output:** R^2_{mean}, $R^2_{\text{mean,DE}}$, R^2_{var}, $R^2_{\text{var,DE}}$
3: **for** $(a, v) \in \texttt{cells}_{\texttt{perturbed}}$ **do**
4: $\quad \mathcal{C}(a, v) \leftarrow \{\ \texttt{cell} \in \texttt{cells}_{\texttt{perturbed}} \mid \texttt{cell}$ received $(a, v)\ \}$.
5: $\quad$ **if** $|\mathcal{C}(a, v)| > 30$ **then**
6: $\quad\quad$ **for** $y \in \texttt{cells}_{\texttt{source}}$ **do**
7: $\quad\quad\quad \hat{\mu}(y, a, v), \hat{\sigma}^2(y, a, v) \leftarrow \texttt{autoencoder}(y, a, v)$.
8: $\quad\quad$ **end for**
9: $\quad\quad$ Compute sample means $\overline{\hat{\mu}}(a, v)$ and $\overline{\hat{\sigma}^2}(a, v)$ of predictions.
10: $\quad\quad$ Compute entrywise variance $\overline{\omega}(a, v)$ of $\{\hat{\mu}(y, a, v)\}_{y \in \texttt{cells}_{\texttt{source}}}$.
11: $\quad\quad$ Compute sample mean $\overline{y}(a, v)$ and sample variance $s^2(a, v)$ of $\mathcal{C}(a, v)$.
12: $\quad\quad$ Compute $R^2_{\text{mean}}(a, v)$ for $\overline{y}(a, v)$ and $\overline{\hat{\mu}}(a, v)$ using all genes.
13: $\quad\quad$ Compute $R^2_{\text{mean,DE}}(a, v)$ for $\overline{y}(a, v)$ and $\overline{\hat{\mu}}(a, v)$ using top DE genes.
14: $\quad\quad$ Compute $R^2_{\text{var}}(a, v)$ for $s^2(a, v)$ and $\overline{\hat{\sigma}^2}(a, v) + \overline{\omega}(a, v)$ using all genes.
15: $\quad\quad$ Compute $R^2_{\text{var,DE}}(a, v)$ for $s^2(a, v)$ and $\overline{\hat{\sigma}^2}(a, v) + \overline{\omega}(a, v)$ using top DE genes.
16: $\quad$ **end if**
17: **end for**
18: Compute sample means R^2_{mean}, $R^2_{\text{mean,DE}}$, R^2_{var}, and $R^2_{\text{var,DE}}$ over all $(a, v) \in \texttt{cells}_{\texttt{perturbed}}$.

Hyperparameters

Table 3.2 lists the hyperparameters of CPA used in experiments.

Table 3.2: CPA hyperparameters used in experiments.

Group	**Hyperparameter**	**Value**
General	Embedding dimension	256
	Batch size	64
	Learning rate decay in epochs	25
Nonlinear scalers	Hidden neurons per layer	32
	Hidden layers	1
	Learning rate	2.2×10^{-3}
	Weight decay	3.9×10^{-8}
Encoder and decoder	Hidden neurons per layer	256
	Hidden layers	8
	Learning rate	1.3×10^{-3}
	Weight decay	1.7×10^{-5}
Adversary	Hidden neurons per layer	512
	Hidden layers	3
	Regularization strength	2000
	Gradient penalty	2.4×10^{-2}
	Learning rate	4.2×10^{-3}
	Weight decay	10^{-7}
	Number of learning steps	3

Chapter 4

Impacts of COVID-19 Interventions

4.1 Introduction

As the world continues to battle coronavirus disease 2019 (COVID-19), growing evidence indicates that the pandemic is exacerbating inequalities in the US [15, 109, 60, 137, 78, 26, 81, 119, 146]. Many studies have focused on racial and ethnic disparities in health outcomes [15, 109, 60, 137, 78, 26, 81]. For example, non-Hispanic African American patients were more than twice likely to be hospitalized than non-Hispanic white patients in a large health care system in California [15]. Overrepresentation of Non-Hispanic black patients among COVID-19 hospitalizations has also been found in Louisiana [109] and Georgia [60]. Not only are racial and ethnic minorities at increased risk of comorbidities that are associated with severe illness [74, 130, 54, 7], but also they are disadvantaged by structural factors such as residential segregation and employment in essential services [137].

In addition to minority status, COVID-19 case and death rates are often higher in urban counties that rank lower in socioeconomic status, housing, and transportation [78]. According to studies of New York City, the test positivity rate was high in neighborhoods that were characterized by poverty, big households, and a large non-Hispanic black or immigrant population [26, 119]. Similar conclusions were drawn about Massachusetts [81]. The high positivity rate was partially explained by insufficient testing that was available to people in poverty and minority groups [89].

Moreover, poorer areas across the US exhibited less physical distancing [146, 72]. This is particularly disturbing because adequate testing and quarantining have been shown to effectively stop the spread of the disease [50].

The growing body of evidence that the COVID-19 pandemic is worsening inequalities in the US stresses the importance of understanding how public policy influences different communities [8]. Socioeconomically disadvantaged regions not only tend to have higher COVID-19 death rates but also are less resilient to economic distress [137, 78, 59, 91]. The latter, which can be manifested in unemployment, may further lead to deaths of despair from suicide, drug overdose, and alcoholism [33, 126]. This link between economic hardship and deaths of despair suggests a possible trade-off between recession-related deaths and COVID-19 deaths.

Although obtaining a clear definition of vulnerability amid the pandemic remains elusive [128, 40], it is clear that certain features correlate with bad outcomes. Our analysis of census, mobility, and COVID-19 data of the US confirms disparities in both COVID-19 deaths and deaths of despair. The strongest predictors for the regional COVID-19 death rate are income, age, race, and household overcrowding. Moreover, we find that regions with worse health outcomes also tend to have higher unemployment and eviction rates. We further investigate the effects of income and household overcrowding on health and economic outcomes. Our analysis confirms the widely believed trade-off between COVID-19 deaths and economic distress-related deaths as the level of lockdown changes [128, 19, 139, 23]. However, we find that this trade-off only exists among socioeconomically disadvantaged counties. Furthermore, the percent of overcrowded households and the COVID-19 death rate are positively correlated. Although our data analysis is inconclusive on whether the identified effects are causal, we answer this question affirmatively by reproducing similar results using agent-based modeling.

While it is crucial that government interventions reduce inequality during the pandemic, designing good interventions is challenging [8, 128, 19]. First, multiple criteria, such as health and economic impacts, can be used for policy evaluation, which may give conflicting advice [19, 139]. Second, it is often hard to estimate

the causal effects of a single intervention from data [146]. We can consider society fighting a pandemic as a complex system that has time-varying nonlinear interactions. Moreover, multiple interventions are usually at work simultaneously [66, 53]. Third, data only exists for the policies that have been implemented [139]. Finally, granular data needed for definitive conclusions are sometimes scarce and incomplete [9].

In order to overcome the limitations of observational data, we develop an agent-based model that simulates the transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the consequent rise in deaths of despair. The model incorporates key elements including socioeconomic status, age-dependent risks, household transmission, asymptomatic transmission, and hospital capacity. Agent-based modeling enables analysis of causal links between various policies and metrics of interest [6, 149, 98, 49, 52, 121]. We investigate the effects of four nonpharmaceutical interventions (NPIs) on inequality, namely, lockdown, testing along with contact tracing, government subsidization, and housing provision. We use the COVID-19 death number to measure health outcomes and deaths of despair as a proxy for economic consequences. Our model generates a stylized population that comprises socioeconomically disadvantaged and privileged people, referred to as poor and rich, respectively, for brevity. As shown with US data, we find that the trade-off between COVID-19 deaths and deaths of despair, hinging on the lockdown level, only exists in the poor community. While subsidization narrows the socioeconomic gap in deaths of despair, the combination of testing, contact tracing, and home isolation alone is effective at reducing disparities in both types of death. Similar to our data findings, our model also suggests a strong link between household overcrowding and the COVID-19 infection rate, which we quantify with mathematical analysis.

Our simulation not only reflects patterns observed in US data but also yields new insights that fill in the gaps of our data analysis. Our findings demonstrate the importance of targeted intervention design to relieve both health-related and economic pressure on socioeconomically disadvantaged populations. Our model suggests a moderate lockdown, adequate testing combined with contact tracing and home isolation, sufficient targeted subsidies, and mitigation of overcrowding in housing. Our results

contribute to policy modeling and evaluation for reducing inequalities during a pandemic. The paper focuses on the US, but our approach and results can be extended to other regions in the world.

4.2 Main Results

4.2.1 Data Analysis

We start by building a decision tree to identify the strongest predictors for the regional COVID-19 death rate. A decision tree is a predictive model that sequentially partitions an input dataset into subsets so that prediction accuracy improves after each split [25]. Decision tree learning provides a natural method of feature selection by quantifying the contribution of each feature to the prediction task [77]. We use census [135, 39] and eviction data [103] from 2019, and COVID-19 death data [102] from 2020 in New York City by ZIP Code Tabulation Area (ZCTA). The census data contains many factors including household overcrowding, the percent of 65-and-older population, the percent of home-based workers, commuting, health insurance coverage, median income, and race.

Fig. 4-1a shows a pruned tree that is fitted to the ZCTA-level data (see the complete decision tree in Fig. 4-5 of Section 4.5). The top scatterplot contains all ZCTAs in the dataset. Income is identified as the feature that best splits the set with a threshold at US$122,200. The percent of 65-and-older population is the best variable to further split the lower-income group (at 17.85%), whereas the percent of household overcrowding is chosen to divide the higher-income group (at 3.72%). The decision tree is built iteratively this way. Although our goal with the dataset is to evaluate feature importance rather than predict the death rate, the decision tree sheds light on the link between regional characteristics and local health outcomes. High COVID-19 death rates are often associated with low income, a large population of seniors and racial minorities, lack of health insurance, high eviction rates, household overcrowding, commuting, and uncommonness of working from home. An excep-

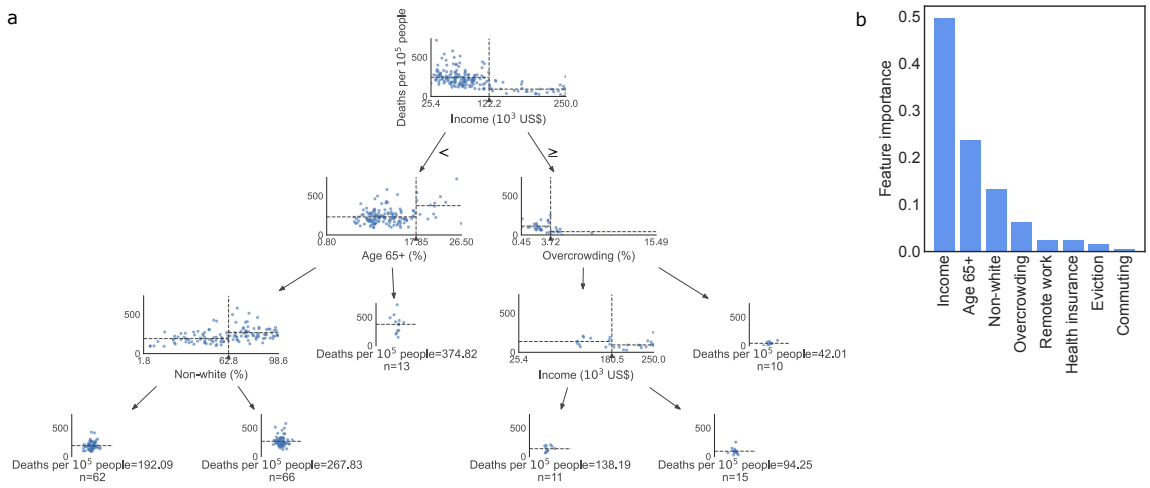

Figure 4-1: Regional features associated with local COVID-19 death rates. a, We build a decision tree that predicts the COVID-19 death rate of New York City by ZCTA. We show a pruned tree here to illustrate the method and provide the full tree in Fig. 4-5 of Section 4.5. The x and y-axes of each scatterplot are the feature used for the split and the number of deaths per $100,000$ people, respectively. ZCTAs are divided into two subsets at the vertical lines so that the death rates are close to the average (marked by horizontal lines) within each group. b, We compute the importance of a feature in the decision tree as the normalized total reduction of the mean squared error that is attributable to the feature.

tion to this pattern is the first appearance of overcrowding in the decision tree as shown in Fig. 4-1a. Surprisingly, the ZCTAs with more household overcrowding had lower death rates. It turns out that these ZCTAs are mostly in Lower and Midtown Manhattan where single young professionals with high salaries tend to live.

We also compare the best and worst segments in the decision tree and find economic inequality in addition to health disparities. Not only did the worst segment have a higher unemployment rate (3.03%) than the best one (2.88%) in 2019, but the former group also had a steeper increase (5.69%) in 2020 than the best segment (4.22%). The 2020 unemployment rates are projected at the ZCTA level by calculating the percent change in unemployment of the county containing the ZCTA and applying this change to the ZCTA level data from 2019. The unemployment gap coincides with the differential eviction rates, which are 0.37% and 0.30% for the worst and the best segments, respectively.

Having learned a decision tree, we then compute the importance of a feature as the

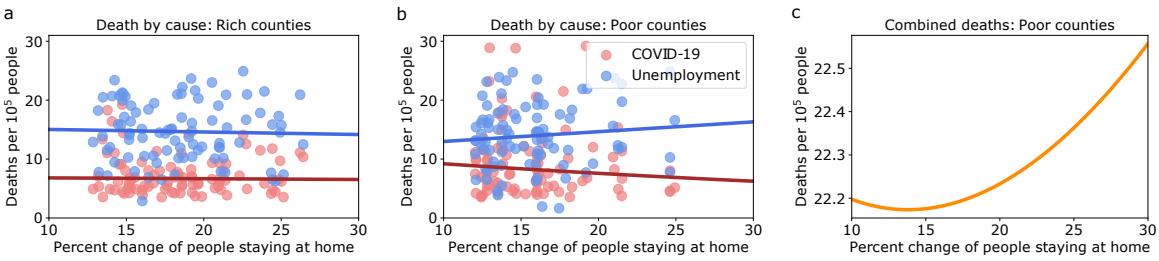

Figure 4-2: Lockdown and social distancing measures that are meant to curb the spread of COVID-19 can exacerbate inequalities. We compare the richest (a) and poorest (b) counties in the US as measured by median income. a, Affluent counties are resilient to the economic shock of lockdown and social distancing measures. b, In contrast, poor counties face the dilemma of whether to die from COVID-19 infection or economic distress. c, Combining estimates from both regression reveals the health and economic trade-off for poor counties.

normalized total reduction of the mean squared error in estimating the COVID-19 death rate of a ZCTA that is attributable to the feature. As shown in Fig. 4-1b, the highest-scoring features are income (0.50), the percent of 65-and-older population (0.24), the percent of non-white population (0.13), and household overcrowding (0.06).

We further investigate the effects of income on regional health and economic outcomes. We compare the poorest and richest counties in the US as measured by median income and find that the widely believed health and economic trade-offs of lockdowns only exist in poor counties (Fig. 4-2). The annual median personal income is less than US$70,000 for the poorest counties, in contrast to above US$80,000 for the richest counties. For this analysis, we combine datasets that measure median income, the unemployment rate, the size of the labor force, the percent change from baseline of people staying at home (as a measure of lockdown severity), and COVID-19 death counts [135, 134, 61, 67].

One puzzle presented by the data is that the level of lockdown appears to be positively correlated with the COVID-19 death rate. Our hypothesized reason is that locations with the most severe outbreaks responded with the most drastic measures. After accounting for disease progression and reporting delays, we observe that stricter policies correspond to lower death rates in poor counties whereas the correlation is weak for rich counties, with the latter possibly due to residual effects from the first wave of COVID-19 (Fig. 4-2a,b). Specifically, we perform linear regression of

the logarithmic transformation of the COVID-19 death rate on the mobility change, delaying the death data by 62 days. Fig. 4-2a,b indicate that there is indeed a damping effect of lockdown and social distancing measures on COVID-19 transmission, which is consistent with conclusions in [66, 16, 105, 58].

In order to compare economic impacts with health outcomes, we project excess deaths caused by economic downturns. Prior work has shown that unemployment increases an individual's mortality hazard by at least 73% [127]. Although the aggregate mortality effects of economic stagnation are open to question, the increased hazard of death associated with individual joblessness has been well established [127, 115, 117, 23]. Using the individual risk inferred in [127], we project the one-year death count attributable to the pandemic-related unemployment shock (Fig. 4-2a,b). Specifically, we estimate the total number of newly unemployed workers in each county using the size of the labor force and the increase in the unemployment rate in 2020 compared to 2019. We then use the all-cause mortality rate from 2019 of each county to calculate the mortality rate of the newly unemployed workers. Finally, we perform linear regression of the projected death rate associated with unemployment on the mobility change. As shown in Fig. 4-2a,b, the unemployment shock affects poor counties more than the rich ones. One explanation is that the reduction in mobility was significantly more in wealthier areas than poorer areas during the pandemic [146], which indicates that the affluent can weather the economic repercussions of lockdowns partially because their jobs allow for flexibility in terms of working remotely. Prior work has drawn similar conclusions that excess mortality is disproportionately high in disadvantaged groups such as African Americans and people with low educational attainment [94, 23].

Fig. 4-2a,b suggest that the widely believed health and economic trade-offs of lockdowns only exist in poor counties. Fig. 4-2c illustrates this trade-off by summing regression estimates of COVID-19 deaths and projected excess deaths attributable to unemployment. Our findings confirm marked differences in the way that social distancing and lockdown measures impact different groups.

We also explore the association between household overcrowding and regional

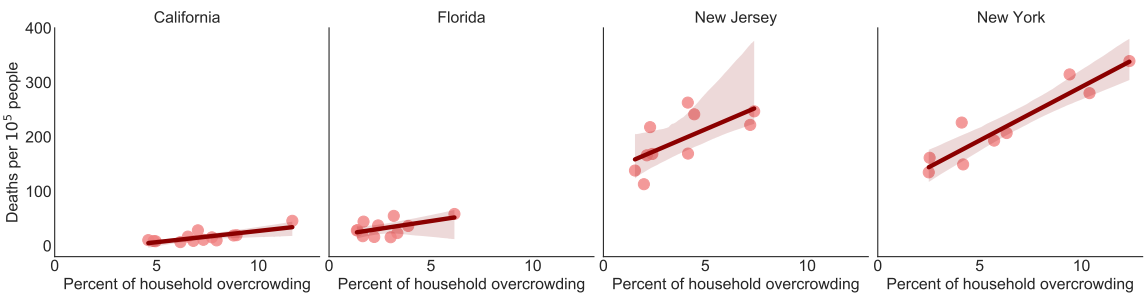

Figure 4-3: The COVID-19 death rate is positively correlated with household overcrowding in urban counties. California, Florida, New Jersey, and New York are the largest four states for the number of counties of which at least 95% of the population live in urban areas. For each state, the solid line and the shaded area represent robust linear regression that downweights outliers with a 95% confidence interval.

health outcomes. Household overcrowding is the condition where there is more than one person per room [132], which may accelerate the spread of respiratory diseases such as COVID-19. We use the Comprehensive Housing Affordability Strategy (CHAS) data prepared by the US Census Bureau for the 2013–2017 period [132]. We focus on the largest four states for the number of urban counties, which are California, Florida, New Jersey, and New York. A county is urban if at least 95% of the population live in urban areas. The rurality data is published by the US Census Bureau for the year 2010 [131]. We restrict the death data [67] to the end of July 2020 to take into account roughly the first six months since the first recorded US case. The qualitative results remain the same as the time window considered changes. Fig. 4-3 indicates a positive correlation between the percent of household overcrowding and the COVID-19 death rate. However, data of rural areas appears particularly noisy (Fig. 4-6 of Section 4.5). This may be explained by several reasons including low population density, large regional variations in infection patterns, and disease outbreaks at different times.

Our findings imply an underlying mechanism at play that causes worse health and economic outcomes for poorer communities. Although our data analysis is inconclusive on whether the identified effects are causal, we answer this question affirmatively by reproducing similar results using agent-based modeling.

4.2.2 Model

We develop an agent-based model that simulates the transmission of SARS-CoV-2 and the consequent rise in deaths of despair. The model takes into account key factors such as socioeconomic status, age-dependent risks, household transmission, asymptomatic transmission, and hospital capacity. We examine the effects of four NPIs on inequality, which are lockdown, testing along with contact tracing, government subsidization, and housing provision. We overview the model in this section, providing details in Section 4.4.

The model initializes a population where each individual has their own attributes that influence their state transitions during simulation. We sample each individual's age from the distribution as specified by the US Census Bureau's 2019 national population estimates [133]. Our stylized model considers people under age 20 as students, those aged 20 to 69 years as workers, and people aged over 69 as retirees. Moreover, everyone is economically active at the start of a simulation. An active individual's output is the sum of the personal output and the connection output, the latter being a measure of the benefits of staying connected to society. Once infected, an individual progresses stochastically from asymptomatic or presymptomatic, to symptomatic, hospitalized, admitted to the ICU, and deceased, with the possibility of recovery at any stage if not deceased (Fig. 4-7a of Section 4.5). Epidemiological parameters and their sources [37, 129, 51, 87, 140] are in Fig. 4.1 of Section 4.5. An individual is economically inactive during hospitalization and at death. Moreover, an individual loses connection output while in quarantine or staying home (Fig. 4-7a of Section 4.5). Taking into account factors that vary across communities such as the comorbidity rate and health care quality [15, 137, 78], we assume that a small fraction of the population are vulnerable to severe illness, exclusive to the poor community. Once infected, vulnerable people are more likely to experience worsening symptoms than an average person.

We incorporate in our model random graphs to simulate virus transmission and economic activities. In consideration of the high transmission rate in households [37,

92], we construct a collection of complete graphs to represent households where any pair of members in the same household are connected. To capture socioeconomic disparities, we assume that 90% of the population are poor and the rest are rich in expectation. A rich person is characterized by a high output and a small household size. In addition, we overlay the household network with an economic network that represent economic activities which rely on in-person contact (Fig. 4-7b of Section 4.5). We generate economic networks using the Watts–Strogatz random graph [145], a classic model that produces the small-world phenomenon as observed in many real-world networks.

Our model considers dynamics at both household and aggregate levels, which include deaths of despair, recession, and undertreatment. We take into consideration deaths of despair that are linked to financial stressors. Specifically, the probability that an individual dies from despair is a function that decreases with per capita output in the household. At the aggregate level, with government subsidies taken into account, a drop in the total output leads to more workers becoming economically inactive. In addition, our model incorporates the scenario in which hospitals are overwhelmed and poor patients are undertreated. Undertreatment increases the chance of deterioration in patients.

4.2.3 Impacts of NPIs on Inequality

It has been widely accepted by now that there is a trade-off between saving lives from the pandemic and saving lives from recession. What has been less scrutinized, however, is how this trade-off varies in different communities and under various policies [19, 139, 23]. As we have observed in US data, poorer counties not only have had more COVID-19 deaths but also will see more recession-induced deaths. We investigate the effects of four NPIs on inequality, which are lockdown, testing along with contact tracing, government subsidization, and housing provision. Our model suggests that, for most NPIs considered, the poor community suffers significantly more than the rich counterpart in terms of both types of death.

Unless stated otherwise, we simulate the dynamics within the population for 180

days, initializing the percent of infections to 0.1%. We assume that retirees stay at home in all simulations, as this policy has been commonly recommended for reducing COVID-19 hospitalizations and deaths [2]. In order to unravel the causal effects of NPIs on inequality, we design experiments so that only one NPI is altered at a time. The baseline setting comprises a lockdown starting on the sixth day at the 0.4 level, a daily testing rate of 0.00145 (0.145% of the population), contact tracing with a success rate of 0.7, need-based subsidies of 0.1, and maximum sizes of rich and poor households at 3 and 5, respectively. We are interested in the potentials of testing, contact tracing, and home isolation, so we set aside lockdown and subsidies while varying the testing rate.

For a lockdown level of $0 \leq \psi \leq 1$, each worker stays at home with probability ψ, independently of the others. Fig. 4-4a shows that the trade-off between COVID-19 deaths and deaths of despair, dependent on the lockdown level, is specific to the poor community, which is consistent with our conclusion from US data (Fig. 4-2). Tightening lockdown from mild ($\psi = 0$, only retirees staying at home) to moderate ($\psi = 0.4$) significantly reduces COVID-19 deaths for both groups. With further lockdown restrictions, marginal health benefits decline, while more poor people die from despair. By contrast, the rich community has almost no deaths of despair and only benefits from a strict lockdown.

Our model uses reverse transcription polymerase chain reaction (RT-PCR) tests with 90% sensitivity and 100% specificity [138]. Given a testing rate, we conduct random testing among susceptible, asymptomatic, and presymptomatic individuals. Once someone tests positive, the person will self-isolate at home until recovery. The person's household members and other contacts will subsequently be prioritized in testing, with the latter being found by contact tracing with a probability of 0.7. Fig. 4-4b suggests that, even without any other NPI, the combination of testing, contact tracing, and home isolation alone is effective at reducing disparities in both types of death. Our findings corroborate the conclusion in [38] that increased testing and contact tracing capacity enables reopening at a larger scale.

We consider government subsidies that are given to anyone in need regardless

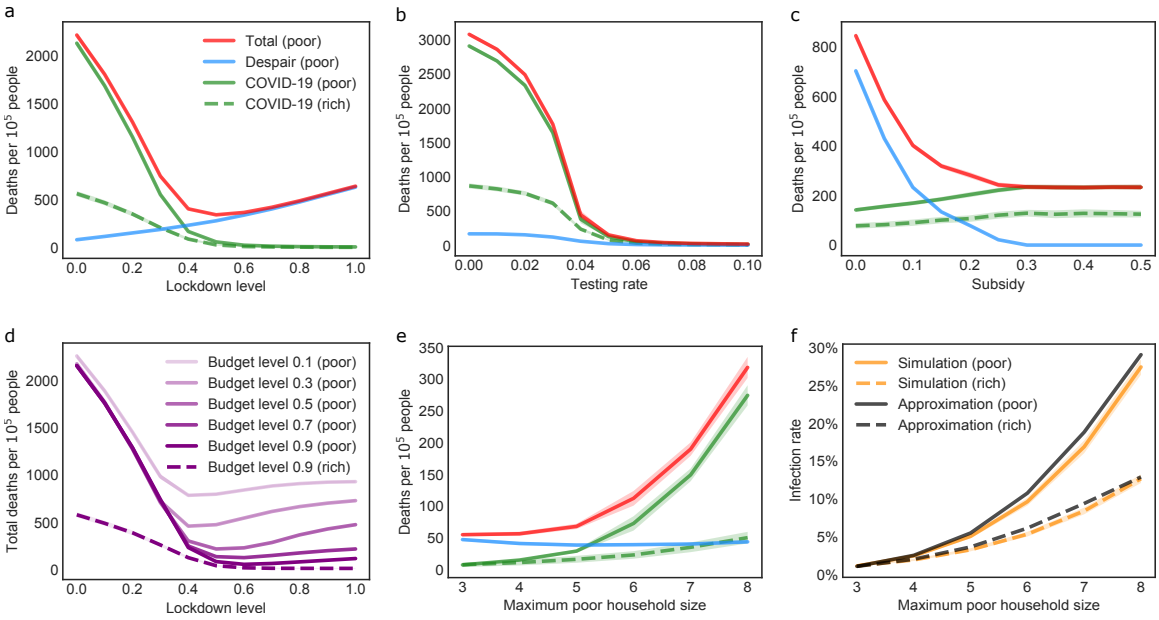

Figure 4-4: Impacts of COVID-19 NPIs on socioeconomic inequality. The fatality rate is calculated within each socioeconomic group. Since the rate of death of despair is close to zero for the rich community, we only show COVID-19 deaths for this group. a, The trade-off between COVID-19 deaths and deaths of despair only exists in the poor community. b, The combination of testing and contact tracing alone is sufficient for eliminating socioeconomic disparities in both types of death. c, Increasing subsidies effectively reduces the gap in deaths of despair. d, For the strategy of prioritizing the neediest people for subsidies, a larger budget narrows disparities in the total death rate and enables stricter lockdown before economic consequences exceed marginal health benefits. Since the rate of death of despair is almost the same for the rich community at all budget levels, we only show this group's results at a budget level of 0.9. e, Household overcrowding exacerbates COVID-19 in the poor community. f, The effect of household overcrowding can be explained by mean-field approximation. Curves and shades are the averages and the standard deviations of 100 trials, respectively.

of socioeconomic status. On each day of simulation, the model looks for and gives money to low-output people who may die from despair. The subsidy is measured as a fraction of an economically active poor individual's personal output. Fig. 4-4c indicates that need-based subsidies no less than 0.3 effectively eliminate the gap in deaths of despair. We also explore the efficacy of greedy subsidization subject to budget constraints. Specifically, given a budget, individuals with the lowest outputs are the ones that are most likely to be impacted by economic volatility and hence prioritized for payment. The budget level is measured as the fraction of the population that can be supported if each subsidy is 0.3. Fig. 4-4d suggests that increasing the

budget level reduces disparities in the total death rate and enables stricter lockdown before economic consequences exceed marginal health benefits.

We investigate the effects of household overcrowding by varying the maximum size of poor households. The configurations of rich households are kept at a maximum size of three and 10% of the population. For ease of mathematical analysis, lockdown starts at initialization, and simulation runs for 60 days. As shown in Fig. 4-4e, a larger difference in household size leads to higher inequality in COVID-19 deaths. This result confirms the causal link between household overcrowding and the COVID-19 death rate suggested by US data (Fig. 4-3). Inspired by [50], we quantify the dependence of the infection rate on household size using mean-field approximation. We denote the average size of poor and rich households by n_p and n_r, respectively. Let I_0 be the number of infections at initialization. Let I_t^p be the number of newly infected poor individuals at time step t. We define I_t^r similarly. Let η_t be the estimated susceptible fraction of the population at time step t. Let ϵ be the secondary attack rate. We use Φ to represent the power of secondary infections that originate from economic connections. We can derive mean-field approximation by

$$\mathbb{E}\left[I_t^p\right] \approx n_p I_0 \left[\eta_t \Phi \left(1 + \epsilon \eta_t \Phi\right)\right]^t, \quad \mathbb{E}\left[I_t^r\right] \approx \frac{n_r}{n_p} \mathbb{E}\left[I_t^p\right]. \tag{4.1}$$

We provide detailed derivation in Section 4.4. It is noteworthy that the ratio between poor and rich communities' infection rates is almost equal to the ratio of average household size. Fig. 4-4f shows that Eq. 4.1 approximates simulation results well.

Since no definitive conclusions have been drawn about the possible link between intergenerational coresidence and the fatality rate [49, 9], we test the robustness of our results against transmission within multigenerational households by letting household members be in the same age group. All qualitative observations remain the same (Fig. 4-8 of Section 4.5).

4.3 Discussion

Although medical science has advanced by leaps and bounds since a century ago when the 1918 influenza pandemic claimed tens of millions of lives worldwide, many challenges remain in the face of a pandemic respiratory illness [122, 28]. It is crucial that we learn from the past and the present in order to prepare for future pandemics. In this paper, we have focused on modeling and evaluating NPIs during the initial stage of a pandemic, taking into account the specifics of COVID-19. We have investigated the differential causal effects of NPIs on different communities using both US data and agent-based modeling. We have identified a socioeconomic gap in both health and economic measures in most situations. Both our data analysis and our simulations have demonstrated that the widely believed health and economic trade-offs of lockdowns only exist in the socioeconomically disadvantaged population. Moreover, household overcrowding leads to increased rates of infection. We have further shown using mean-field approximation that the ratio between two communities' infection rates is almost equal to the ratio of average household size. Our model has suggested that, even without any other NPI, the combination of testing, contact tracing, and home isolation alone is effective at reducing disparities in COVID-19 and recession-related deaths. Our simulations have also shown the efficacy of targeted subsidies in mitigating the negative economic effects of strict lockdowns, which disproportionately impact disadvantaged groups.

There are several important implications from this work. Our results underline the importance of intervention design in a pandemic as socioeconomically disadvantaged populations bear the brunt of suboptimal policies, which will worsen existing inequalities. Our findings suggest that an effective methodology for confronting COVID-19 is a combination of a moderate lockdown with targeted and sufficient subsidies to mitigate the economic consequences, adequate testing along with contact tracing and home isolation, and easing overcrowding in housing. These measures should be coordinated in order to reduce inequalities under fiscal and logistical constraints. Although we have focused on the US in this paper, our approach and results can be extended

to other regions in the world. For example, the deaths of despair phenomenon in the US can be instead considered as mortality associated with food deprivation in low-income countries. Based on the estimates of [73], 22% of the adult population in Ethiopia, Malawi, Nigeria, and Uganda face severe food insecurity during the pandemic, with higher prevalence in poorer households. Understanding the differential impacts of NPIs on various demographic groups continues to be a pressing issue for low-income countries as COVID-19 vaccine shortages are expected to persist in these regions. Another contribution of this study is to identify factors that make a community more vulnerable to COVID-19 and elucidate their effects under various NPIs, which is closely related to the work on defining vulnerability indices. There is a growing number of vulnerability indices that help guide resource allocation during a pandemic, which, however, may give divergent recommendations [40]. In order to determine an appropriate index, it is essential to understand how policies impact communities differently.

Our study has several limitations. First, the conclusions drawn from our analysis rely on aggregate data at the ZCTA and county levels. Ideally, comprehensive data at the individual or household level which encompass many aspects such as socioeconomic status, medical conditions, and behavior in response to COVID-19 are used to infer the differential causal effects of NPIs on different demographic groups. In practice, such granular data rarely exist due to challenges in collection and privacy. The limitations of the data are partially addressed by our work on agent-based modeling. Second, our simulations are based on a stylized model that captures key elements to the topic studied, including socioeconomic status, age-dependent risks, and household transmission, but leaves out other details. We have chosen to build a medium-sized model in order to obtain qualitative insights. Detailed agent-based models that typically require high-performance computing are needed for drawing quantitative conclusions. Finally, we have only considered lockdowns and testing that are conducted uniformly across the population. In reality, low-income areas across the US have faced obstacles to testing and physical distancing [89, 146, 72]. For this reason, the socioeconomic gap in COVID-19 deaths identified by our model

is a conservative estimate.

There are several interesting directions for future research. One extension is to investigate interventions that are adjusted over time according to feedback and how such adaptive measures affect inequalities. Another interesting avenue of research is exploring how to incentivize safe behavior that can lessen the need for drastic lockdowns. Given the national variations in the vaccine rollout strategy, it is also urgent to understand how to design vaccine programs that reduce inequalities. Additionally, it is important to take into consideration fiscal and logistical constraints for the task of policy evaluation. These questions are not only of much practical relevance to COVID-19 but also fascinating research problems that call for multidisciplinary efforts. Progress towards these goals will have a lasting impact on policy responses to future pandemics.

4.4 Methods

4.4.1 Vulnerable Group

We assume that, on average, 1% of the population are at increased risk for severe illness from SARS-CoV-2, all of whom are poor. We define a vulnerability factor $v > 0$ as the extent to which a vulnerable person is more likely to experience worsening symptoms than the average rate. For example, let μ be the hospitalization rate for people in their 50s who are infected and symptomatic. The probability that someone symptomatic in this age group needs to be hospitalized is $(1 + v)\mu$ if the person is vulnerable. For non-vulnerable individuals, the probability is $(1 - v/99)\mu$. In general, a vulnerable person, once infected, is more likely to move through the disease stages of symptoms, hospital admission, ICU admission, and death by a factor of v than the age-specific average rate.

4.4.2 Networks

Let m_p be the maximum number of people living in a poor household. Similarly, we define m_r as the maximum size of a rich household. Unless stated otherwise, we use $m_p = 5$ and $m_r = 3$ in simulations. Let h_p and h_r be the number of poor and rich households, respectively. To construct a household network, we generate h_p complete graphs where the number of nodes in each complete graph is sampled uniformly at random between 1 and m_p. Similarly, we create h_r rich households. We set $h_r = 34000$ and calculate h_p such that poor people constitute 90% of the population on average. We subsequently use the Watts–Strogatz random graph [145] to generate an economic network on the nodes of the household network. Intuitively, the nodes are first arranged into a ring, and then each node is connected with its k nearest neighbors. Finally, each edge in the economic network is rewired with probability p, independently of other edges. Networks constructed as such are known to exhibit the small-world phenomenon [145]. Unless stated otherwise, we use $k = 20$ and $p = 0.5$ in simulations.

4.4.3 Individual Output

For simplicity, we assume that all the income inequality in society is explained by differences in individual productivity. Other sources of inequality are not addressed. For a poor individual who is economically active, let x_p be the output per economic connection and y_p be the personal output. Thus, the average output of an active poor individual is $O_p = y_p + kx_p$ at initialization. Similarly, we define x_r, y_r, and O_r for rich individuals. Let λ be the rich-to-poor output ratio where $x_r = \lambda x_p$ and $y_r = \lambda y_p$. To capture the wealth inequality in the US [18], we suppose that rich people account for only $0 < \theta \ll 1$ of the population but 45% of the total output. In other words, $\theta O_r = 0.45[\theta O_r + (1 - \theta)O_p]$. For $\theta = 0.1$, solving the equation gives $\lambda = 81/11$. For an economic connection to be counted in an individual's output, we require both persons to be (i) economically active, (ii) not staying at home because of lockdown, and (iii) not in isolation due to COVID-19 symptoms. Assuming that half

of the workers staying at home leads to a drop in the total output by 15%, we can get $y_p = 4kx_p$. Without loss of generality, we let $y_p = 1$ and calculate other variables as discussed.

4.4.4 Deaths of Despair

Taking into consideration deaths caused by financial stress, we suppose that a household with a low per capita output is at increased risk for death of despair. Let O be the per capita output in a household, including subsidies received and excluding members who are hospitalized or deceased. Let $z = O_p - O$ be the difference between the household's per capita output and the average value for poor individuals at initialization. For a despair coefficient $0 < \delta \ll 1$, we define the probability of death of despair by a generalized logistic function $q(z) = \delta[1 + \nu^{z/\omega}]^{-1/\nu}$ where $\omega = kx_p/2$ and $\nu = 0.001$ set the inflection point at $z = kx_p/2$, which equals a poor individual's output loss if economic connections are halved. We use $\delta = 5.5 \times 10^{-5}$ in simulations. Fig. 4-9 of Section 4.5 plots the probability of death of despair with respect to output loss. On each day, we calculate q for every household. Each member of the household then dies from despair on the day with probability q, independently of each other.

4.4.5 Recession

Let O_0 be the total output at initialization. Let O_t be the total output on day t, taking into account subsidies distributed on the day. We define $0 < \beta \ll 1$ as an inactive coefficient. If $O_t < O_0$, then we assume that a worker becomes economically inactive on day t with probability $\beta(1 - O_t/O_0)$, independently of each other. We use $\beta = 0.01$ in simulations.

4.4.6 Undertreatment

If the number of people hospitalized with COVID-19 exceeds the hospital capacity, then hospitalized patients will be at increased risk for severe illness. Let $0 < \gamma \leq 1$ and $\lambda \geq 0$ be the coefficients of hospital capacity and undertreatment effects, respectively.

We denote the population size by N and the current number of COVID-19-associated hospitalizations by H. If $H > \gamma N$, then hospitalized poor patients will be more likely to be admitted to the ICU and possibly die later by a factor of $\lambda[H/(\gamma N) - 1]$ than their age-specific risks. By contrast, we assume that rich patients are not affected by overwhelmed hospitals. We use $\gamma = 0.0025$ and $\lambda = 0.5$ in simulations.

4.4.7 Mean-Field Approximation

We denote the average size of poor and rich households by $n_p = (1 + m_p)/2$ and $n_r = (1 + m_r)/2$, respectively. Since the majority of the population are poor, our approximation first assumes that all households are poor and then considers rich households at the end.

Let d be the average number of days that an infected person is asymptomatic or pre-symptomatic. Since only a small fraction of infections lead to hospitalization and more severe outcomes, we ignore these cases in our approximation. The same as in simulation, we assume that anyone symptomatic quarantines at home. In other words, $d \approx ad_a + (1 - a)d_s$, where a is the asymptomatic infection rate, d_a is the average number of days of illness until recovery for asymptomatic patients, and d_s is the average number of days of illness until symptom onset for symptomatic patients. We define one time step as d days.

Let I_0 be the number of infections at initialization. Let I_t^p be the number of newly infected poor individuals at time step t. We suppose that, each day, an infected person who has no symptoms spreads the virus to any of her connections from a different household with probability $\rho > 0$. Given the high risk of household transmission, we suppose that, once someone is infected, everyone else in the same household is immediately infected. Thus, the effective number of initial infections is $n_p I_0$. We consider that everyone stays at home with probability $1 - \alpha$, independently of any other event. In order for an infected person to infect someone from a different household, both persons need to leave home, which occurs with probability α^2. Under the assumption that each person's economic connections are from different households, every infected individual spreads the disease to $\alpha^2 k\rho d$ economic connections in one

time step on average. Infected connections then immediately infect their household members. Moreover, we suppose that a fraction $0 < \epsilon < 1$ of these new infections further spread the disease to their economic connections and hence their household. Thus, I_1^p is roughly equal to $n_p I_0 \Phi(1 + \epsilon\Phi)$ on average where $\Phi = n_p \alpha^2 k \rho d$. Several assumptions such as immediate household transmission and uniqueness of economic connections' households make our estimated number of infections an overestimation, which becomes more marked as time goes on. We adjust for the error by taking into account the susceptible population that shrinks over time. Let $\eta_t = 1 - \sum_{\tau=0}^{t-1} \mathbb{E}\left[I_\tau^p\right]/N$ be the estimated susceptible fraction of the population at time step t. We then have $\mathbb{E}\left[I_1^p\right] \approx n_p I_0 \eta_1 \Phi(1 + \epsilon\eta_1\Phi)$. By induction on time, we have

$$\mathbb{E}\left[I_t^p\right] \approx n_p I_0 \left[\eta_t \Phi \left(1 + \epsilon \eta_t \Phi\right)\right]^t.$$

Let I_t^r be the number of newly infected rich individuals at time step t. The probability that someone from a rich household gets infected can be approximated by n_r/n_p times the infection rate of poor people. Therefore,

$$\mathbb{E}\left[I_t^r\right] \approx \frac{n_r}{n_p} \mathbb{E}\left[I_t^p\right].$$

4.5 Supplementary Information

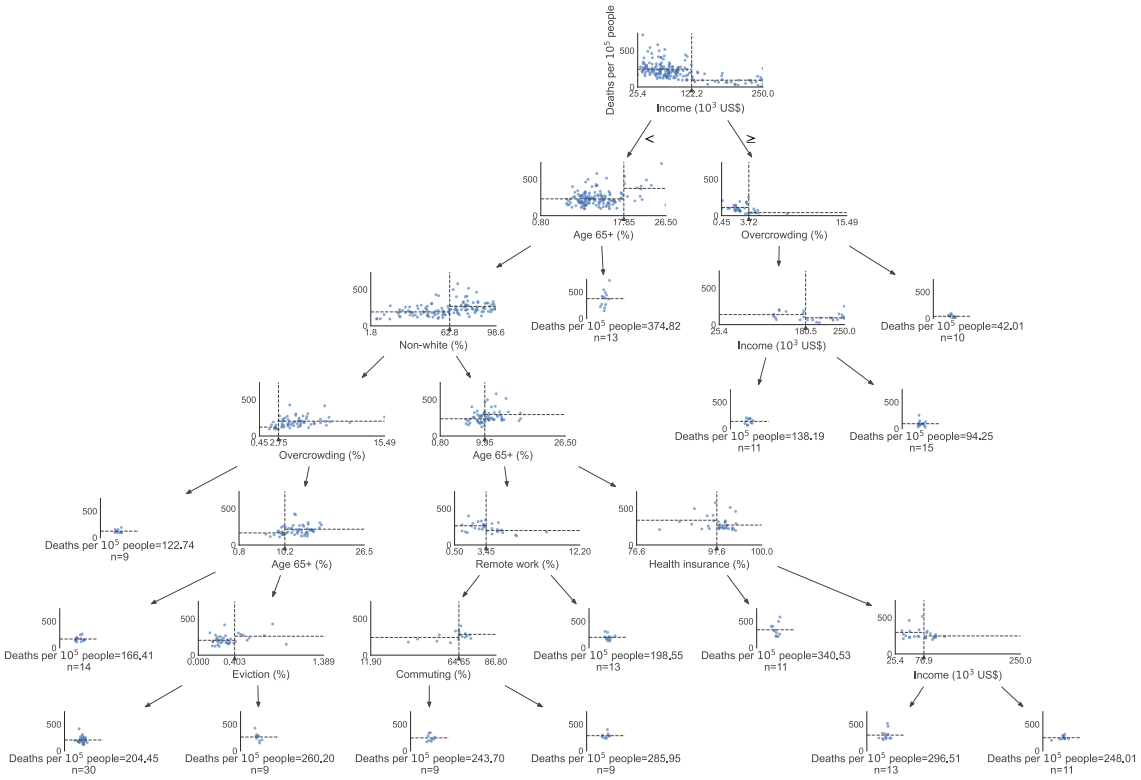

Figure 4-5: A decision tree that predicts the COVID-19 death rate of New York City by ZCTA. The x and y-axes of each scatterplot are the feature used for the split and the number of deaths per $100,000$ people, respectively. ZCTAs are divided into two subsets at the vertical lines so that the death rates are close to the average (marked by horizontal lines) within each group.

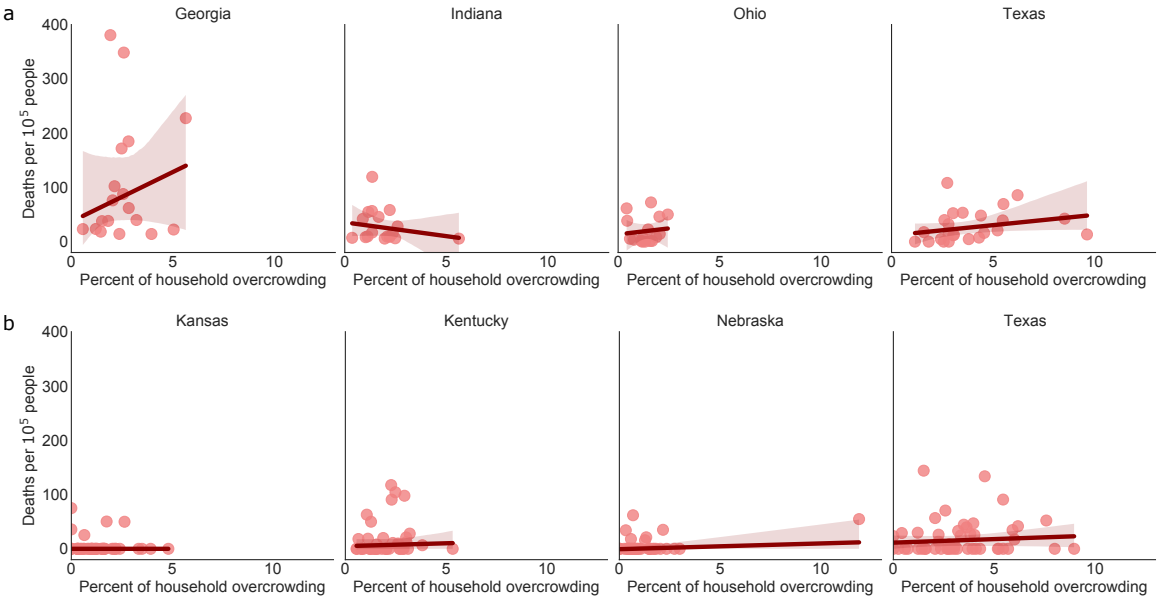

Figure 4-6: The relationship between household overcrowding and the COVID-19 death rate are unclear in rural counties. Potential reasons include low population density, large regional variations in infection patterns, and disease outbreaks at different times. a, The largest four states for the number of counties of which the percent of the population living in rural areas is between 45% and 55%. b, The largest four states for the number of completely rural counties where the whole population live in rural areas. For each state, the solid line and the shaded area represent robust linear regression that downweights outliers with a 95% confidence interval.

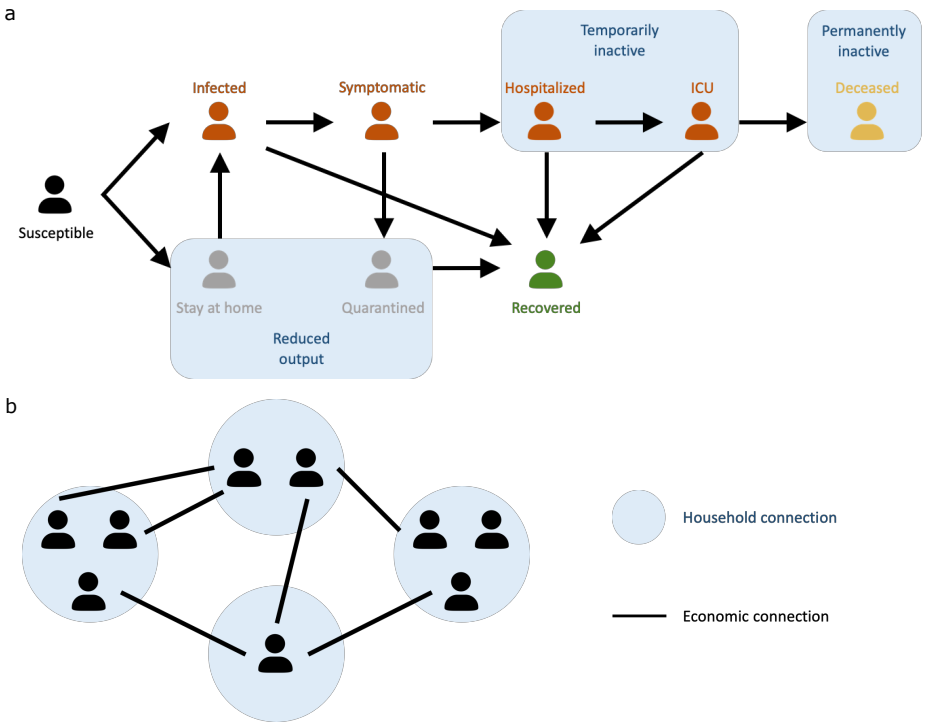

Figure 4-7: Schematic diagrams of the agent-based model. a, Once infected, an individual progresses stochastically from asymptomatic or presymptomatic, to symptomatic, hospitalized, admitted to the ICU, and deceased, with the possibility of recovery at any stage if not deceased. While staying at home, a susceptible individual may still be infected by people in the same household. Once symptomatic, the infected individual quarantines at home until recovery unless hospitalization becomes necessary. An individual is economically inactive during hospitalization and at death. Moreover, an individual loses connection output while in quarantine or staying home. b, Each blue circle corresponds to a complete graph that represents a household. The economic network is generated using the Watts–Strogatz random graph.

Table 4.1: Epidemiological parameter definitions, baseline values, and sources. Time between different stages of infection is sampled uniformly at random from the corresponding intervals listed.

Definition	**Baseline value**	**Source**
Asymptomatic rate	35%	[129]
Probability of hospitalization conditional on symptomatic infection	$\leq$ 9 years: 0.001 10–19 years: 0.003 20–29 years: 0.012 30–39 years: 0.032 40–49 years: 0.049 50–59 years: 0.102 60–69 years: 0.166 70–79 years: 0.243 $\geq$ 80 years: 0.273	[51]
Probability of ICU admission conditional on hospitalization	$\leq$ 39 years: 0.05 40–49 years: 0.063 50–59 years: 0.122 60–69 years: 0.274 70–79 years: 0.432 $\geq$ 80 years: 0.709	[51]
Probability of mortality conditional on ICU admission	$\leq$ 19 years: 0.615 20–39 years: 0.769 40–49 years: 0.748 50–59 years: 0.742 60–69 years: 0.744 70–79 years: 0.747 $\geq$ 80 years: 0.739	[51]
Pre-symptomatic period	2–10 days	[129]
Time from symptom onset to hospitalization	1–12 days	[129]
Time from hospitalization to ICU admission	$\leq$ 14 days	[87]
Time from ICU admission to mortality	$\leq$ 14 days	[87]
Time from symptom onset to recovery	7–28 days	[140]
Probability of infection transmission per contact per day	Household: 0.25 Others: 0.005	[37]

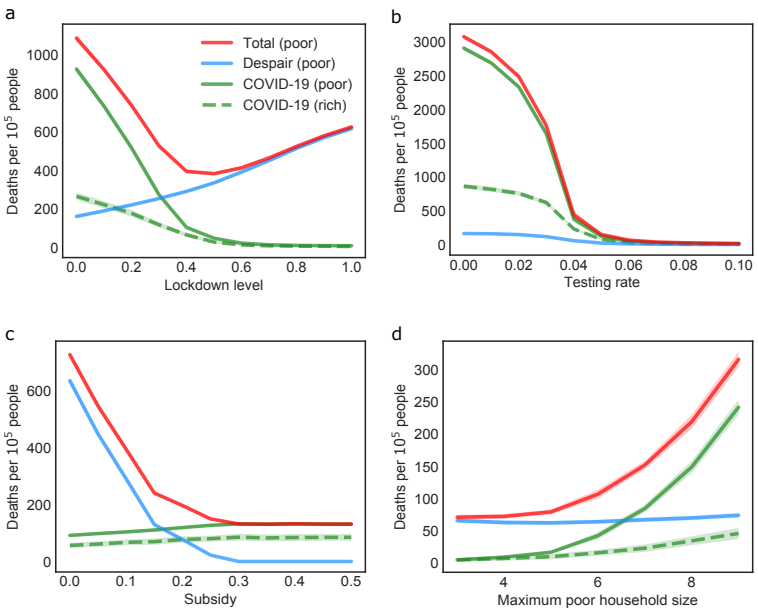

Figure 4-8: Robustness tests for impacts of COVID-19 NPIs on socioeconomic inequality. Each household comprises members from the same age group. All qualitative observations remain the same as those with multigenerational households (Fig. 4-4). The fatality rate is calculated within each socioeconomic group. Since the rate of death of despair is close to zero for the rich community, we only show COVID-19 deaths for this group. a, The trade-off between COVID-19 deaths and deaths of despair only exists in the poor community. b, The combination of testing and contact tracing alone is sufficient for eliminating socioeconomic disparities in both types of death. c, Increasing subsidies effectively reduces the gap in deaths of despair. d, Household overcrowding exacerbates COVID-19 in the poor community. Curves and shades are the averages and the standard deviations of 100 trials, respectively.

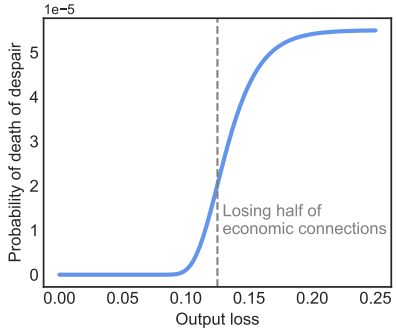

Figure 4-9: Probability of death of despair. The probability that an individual dies from despair increases with per capita output loss in the household.

Bibliography

[1] Yasin Abbasi-Yadkori, Aldo Pacchiano, and My Phan. Regret balancing for bandit and RL model selection. *arXiv preprint arXiv:2006.05491*, 2020.

[2] Daron Acemoglu, Victor Chernozhukov, Iván Werning, and Michael D. Whinston. Optimal targeted lockdowns in a multi-group SIR model. Working Paper 27102, National Bureau of Economic Research, 2020. Available at: `https://www.nber.org/papers/w27102` [Accessed December 30, 2020].

[3] Alekh Agarwal, Miroslav Dudik, Satyen Kale, John Langford, and Robert E. Schapire. Contextual bandit learning with predictable rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2012.

[4] Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38, 2017.

[5] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, 2017.

[6] Alberto Aleta, David Martín-Corral, Ana Pastore y Piontti, Marco Ajelli, Maria Litvinova, Matteo Chinazzi, Natalie E. Dean, M. Elizabeth Halloran, Ira M. Longini Jr, Stefano Merler, Alex Pentland, Alessandro Vespignani, Esteban Moro, and Yamir Moreno. Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nature Human Behaviour*, 4(9):964–971, 2020.

[7] Kamyar Arasteh. Prevalence of comorbidities and risks associated with COVID-19 among Black and Hispanic populations in New York City: An examination of the 2018 New York City community health survey. *Journal of Racial and Ethnic Health Disparities*, pages 1–7, 2020.

[8] Patrick J. Arena, Monica Malta, Anne W. Rimoin, and Steffanie A. Strathdee. Race, COVID-19 and deaths of despair. *EClinicalMedicine*, 25:100485, 2020.

[9] Bruno Arpino, Valeria Bordone, and Marta Pasqualini. No clear association emerges between intergenerational relationships and COVID-19 fatality rates from macro-level analyses. *Proceedings of the National Academy of Sciences*, 117(32):19116–19121, 2020.

[10] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Conference on Learning Theory*, pages 13–26, Haifa, Israel, 27–29 Jun 2010.

[11] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[12] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *IEEE 36th Annual Foundations of Computer Science*, pages 322–331, 1995.

[13] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

[14] Yishai Avior, Ido Sagi, and Nissim Benvenisty. Pluripotent stem cells in disease modelling and drug discovery. *Nature Reviews Molecular Cell Biology*, 17(3):170–182, 2016.

[15] Kristen M. J. Azar, Zijun Shen, Robert J. Romanelli, Stephen H. Lockhart, Kelly Smits, Sarah Robinson, Stephanie Brown, and Alice R. Pressman. Disparities in outcomes among COVID-19 patients in a large health care system in California. *Health Affairs*, 39(7):1–10, 2020.

[16] Hamada S. Badr, Hongru Du, Maximilian Marshall, Ensheng Dong, Marietta M. Squire, and Lauren M. Gardner. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(11):1247–1254, 2020.

[17] José Gabriel Barcia Durán, Raphaël Lis, and Shahin Rafii. Haematopoietic stem cell reprogramming and the hope for a universal blood product. *FEBS Letters*, 593(23):3253–3265, 2019.

[18] Michael Batty, Jesse Bricker, Joseph Briggs, Sarah Friedman, Danielle Nemschoff, Eric Nielsen, Kamila Sommer, and Alice Henriques Volz. *The Distributional Financial Accounts of the United States*. University of Chicago Press, March 2020.

[19] Itai Bavli, Brent Sutton, and Sandro Galea. Harms of public health interventions against covid-19 must not be ignored. *BMJ*, 371:m4074, 2020.

[20] Donald A. Berry, Robert W. Chen, Alan Zame, David C. Heath, and Larry A. Shepp. Bandit problems with infinitely many arms. *Ann. Statist.*, 25(5):2103–2116, 1997.

[21] Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

[22] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2011.

[23] Francesco Bianchi, Giada Bianchi, and Dongho Song. The long-term impact of the COVID-19 unemployment shock on life expectancy and mortality rates. Working Paper 28304, National Bureau of Economic Research, 2020. Available at: `https://www.nber.org/papers/w28304` [Accessed March 15, 2021].

[24] Jonathan Binas, Leonie Luginbuehl, and Yoshua Bengio. Reinforcement learning for sustainable agriculture. *CCAI Workshop at the 36th International Conference on Machine Learning*, 2019.

[25] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[26] George J. Borjas. Demographic determinants of testing incidence and Covid-19 infections in New York City neighborhoods. *Covid Economics, Vetted and Real-Time Papers*, 3:12–39, 2020.

[27] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[28] Aisha Bradshaw. Lessons from the past. *Nature Human Behaviour*, 4(5):448, 2020.

[29] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In Ricard Gavaldà, Gábor Lugosi, Thomas Zeugmann, and Sandra Zilles, editors, *Proceedings of the 20th International Conference on Algorithmic Learning Theory*, pages 23–37, Porto, Portugal, 3–5 Oct 2009. Springer Berlin Heidelberg.

[30] Yosef Buganim, Dina A. Faddah, and Rudolf Jaenisch. Mechanisms and models of somatic cell reprogramming. *Nature Reviews Genetics*, 14(6):427–439, 2013.

[31] Patrick Cahan, Hu Li, Samantha A. Morris, Edroaldo Lummertz da Rocha, George Q. Daley, and James J. Collins. CellNet: Network biology applied to stem cell engineering. *Cell*, 158(4):903–915, 2014.

[32] Alexandra Carpentier and Michal Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*, pages 1133–1141, 2015.

[33] Anne Case and Angus Deaton. *Deaths of Despair and the Future of Capitalism*. Princeton University Press, 2020.

[34] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[35] Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[36] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 151–159, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[37] Hao-Yuan Cheng, Shu-Wan Jian, Ding-Ping Liu, Ta-Chou Ng, Wan-Ting Huang, Hsien-Ho Lin, and the Taiwan COVID-19 Outbreak Investigation Team. Contact tracing assessment of COVID-19 transmission dynamics in Taiwan and risk at different exposure periods before and after symptom onset. *JAMA Internal Medicine*, 180(9):1156–1163, 2020.

[38] Weihsueh A. Chiu, Rebecca Fischer, and Martial L. Ndeffo-Mbah. State-level needs for social distancing and contact tracing to contain COVID-19 in the United States. *Nature Human Behaviour*, 4(10):1080–1090, 2020.

[39] City University of New York. New York City census data: Neighborhood profiles. Data file, 2020.

[40] Columbia Center for Spatial Research and Yale Global Health Partnership. Mapping the new politics of care. `https://newpoliticsofcare.net/`, 2020.

[41] Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, pages 521–529, 2014.

[42] Richard Combes, Alexandre Proutiere, and Alexandre Fauquette. Unimodal bandits with continuous arms: Order-optimal regret without smoothness. *Proc. ACM Meas. Anal. Comput. Syst.*, 4(1), 2020.

[43] Eric W. Cope. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control*, 54(6):1243–1253, 2009.

[44] Edward A. Copelan. Hematopoietic stem-cell transplantation. *New England Journal of Medicine*, 354(17):1813–1826, 2006.

[45] Ana C. D'Alessio, Zi Peng Fan, Katherine J. Wert, Petr Baranov, Malkiel A. Cohen, Janmeet S. Saini, Evan Cohick, Carol Charniga, Daniel Dadon, Nancy M.

Hannett, Michael J. Young, Sally Temple, Rudolf Jaenisch, Tong Ihn Lee, and Richard A. Young. A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Reports*, 5(5):763–775, 2015.

[46] Allard de Wit, Hendrik Boogaard, Davide Fumagalli, Sander Janssen, Rob Knapen, Daniel van Kraalingen, Iwan Supit, Raymond van der Wijngaart, and Kees van Diepen. 25 years of the WOFOST cropping systems model. *Agricultural Systems*, 168:154–167, 2019.

[47] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016.

[48] Gabrielle A. Dotson, Charles W. Ryan, Can Chen, Lindsey Muir, and Indika Rajapakse. Cellular reprogramming: Mathematics meets medicine. *WIREs Mechanisms of Disease*, 13(4):e1515, 2021.

[49] Albert Esteve, Iñaki Permanyer, Diederik Boertien, and James W. Vaupel. National age and coresidence patterns shape COVID-19 vulnerability. *Proceedings of the National Academy of Sciences*, 117(28):16118–16120, 2020.

[50] Sarah C. Fay, Dalton J. Jones, Munther A. Dahleh, and A. E. Hosoi. Simple control for complex pandemics, 2020.

[51] Neil M. Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Ilaria Dorigatti, Han Fu, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Lucy C Okell, Sabine van Elsland, Hayley Thompson, Robert Verity, Erik Volz, Haowei Wang, Yuanrong Wang, Patrick G. T. Walker, Caroline Walters, Peter Winskill, Charles Whittaker, Christl A. Donnelly, Riley, Steven, Azra C. Ghani, and Imperial College COVID-19 Response Team. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Report 9, WHO Collaborating Centre for Infectious Disease Modelling, MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics, Imperial College London, 2020. Available at: `https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/` [Accessed June 4, 2020].

[52] Josh A. Firth, Joel Hellewell, Petra Klepac, Stephen Kissler, CMMID COVID-19 Working Group, Adam J. Kucharski, and Lewis G. Spurgin. Using a real-world network to model localized COVID-19 control strategies. *Nature Medicine*, 26(10):1616–1622, 2020.

[53] Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Juliette T. Unwin, Thomas A. Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W. Eaton, Mélodie Monod, Pablo N. Perez-Guzman, Nora Schmit, Lucia Cilloni, Kylie E. C. Ainslie, Marc Baguelin, Adhiratha Boonyasiri, Olivia Boyd, Lorenzo Cattarino, Laura V. Cooper, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra Djaafara, Ilaria Dorigatti, Sabine L. van Elsland, Richard G. FitzJohn, Katy A. M. Gaythorpe, Lily Geidelberg, Nicholas C. Grassly, William D. Green, Timothy Hallett, Arran Hamlet, Wes Hinsley, Ben Jeffrey, Edward Knock, Daniel J. Laydon, Gemma Nedjati-Gilani, Pierre Nouvellet, Kris V. Parag, Igor Siveroni, Hayley A. Thompson, Robert Verity, Erik Volz, Caroline E. Walters, Haowei Wang, Yuanrong Wang, Oliver J. Watson, Peter Winskill, Xiaoyue Xi, Patrick G. T. Walker, Azra C. Ghani, Christl A. Donnelly, Steven Riley, Michaela A. C. Vollmer, Neil M. Ferguson, Lucy C. Okell, Samir Bhatt, and Imperial College COVID-19 Response Team. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):251–261, 2020.

[54] Katherine M. Flegal, Deanna Kruszon-Moran, Margaret D. Carroll, Cheryl D. Fryar, and Cynthia L. Ogden. Trends in obesity among adults in the United States, 2005 to 2014. *JAMA*, 315(21):2284–2291, 2016.

[55] Dylan J. Foster, Satyen Kale, Mehryar Mohri, and Karthik Sridharan. Parameter-free online learning via model selection. In *Advances in Neural Information Processing Systems*, pages 6020–6030, 2017.

[56] Dylan J. Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 14741–14752, 2019.

[57] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.

[58] Edward L. Glaeser, Caitlin Gorback, and Stephen J. Redding. How much does COVID-19 increase with mobility? Evidence from New York and four other U.S. cities. Working Paper 27519, National Bureau of Economic Research, 2020. Available at: `https://www.nber.org/papers/w27519` [Accessed March 15, 2021].

[59] Dana A. Glei, Noreen Goldman, and Maxine Weinstein. A growing socioeconomic divide: Effects of the Great Recession on perceived economic distress in the United States. *PLoS One*, 14(4):e0214947, 2019.

[60] Jeremy A. W. Gold, Karen K. Wong, Christine M. Szablewski, Priti R. Patel, John Rossow, Juliana da Silva, Pavithra Natarajan, Sapna Bamrah Morris,

Robyn Neblett Fanfair, Jessica Rogers-Brown, Beau B. Bruce, Sean D. Browning, Alfonso C. Hernandez-Romieu, Nathan W. Furukawa, Mohleen Kang, Evans Mary E., Nadine Oosmanally, Melissa Tobin-D'Angelo, Cherie Drenzek, David J. Murphy, Julie Hollberg, James M. Blum, Robert Jansen, David W. Wright, William M. Sewell III, Jack D. Owens, Benjamin Lefkove, Frank W. Brown, Deron C. Burton, Timothy M. Uyeki, Stephanie R. Bialek, and Brendan R. Jackson. Characteristics and clinical outcomes of adult patients hospitalized with COVID-19 — Georgia, March 2020. *Morbidity and Mortality Weekly Report*, 69(18):545–550, 2020.

[61] Google LLC. Google COVID-19 community mobility reports. Data file, 2020.

[62] Thomas Graf and Tariq Enver. Forcing cells to change lineages. *Nature*, 462(7273):587–594, 2009.

[63] Marica Grskovic, Ashkan Javaherian, Berta Strulovici, and George Q. Daley. Induced pluripotent stem cells — opportunities for disease modelling and drug discovery. *Nature Reviews Drug Discovery*, 10(12):915–929, 2011.

[64] Xiaoping Han, Ziming Zhou, Lijiang Fei, Huiyu Sun, Renying Wang, Yao Chen, Haide Chen, Jingjing Wang, Huanna Tang, Wenhao Ge, Yincong Zhou, Fang Ye, Mengmeng Jiang, Junqing Wu, Yanyu Xiao, Xiaoning Jia, Tingyue Zhang, Xiaojie Ma, Qi Zhang, Xueli Bai, Shujing Lai, Chengxuan Yu, Lijun Zhu, Rui Lin, Yuchi Gao, Min Wang, Yiqing Wu, Jianming Zhang, Renya Zhan, Saiyong Zhu, Hailan Hu, Changchun Wang, Ming Chen, He Huang, Tingbo Liang, Jianghua Chen, Weilin Wang, Dan Zhang, and Guoji Guo. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808):303–309, 2020.

[65] Jacob H. Hanna, Krishanu Saha, and Rudolf Jaenisch. Pluripotency and cellular reprogramming: Facts, hypotheses, unresolved issues. *Cell*, 143(4):508–525, 2010.

[66] Nils Haug, Lukas Geyrhofer, Alessandro Londei, Elma Dervic, Amélie Desvars-Larrive, Vittorio Loreto, Beate Pinior, Stefan Thurner, and Peter Klimek. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nature Human Behaviour*, 4(12):1303–1312, 2020.

[67] Hopkins Population Center. COVID-19 SES data hub. Data file, 2020.

[68] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), 2019.

[69] Masaki Ieda, Ji-Dong Fu, Paul Delgado-Olguin, Vasanth Vedantham, Yohei Hayashi, Benoit G. Bruneau, and Deepak Srivastava. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, 142(3):375–386, 2010.

[70] Rudolf Jaenisch and Richard Young. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell*, 132(4):567–582, 2008.

[71] Diego Adhemar Jaitin, Assaf Weiner, Ido Yofe, David Lara-Astiaso, Hadas Keren-Shaul, Eyal David, Tomer Meir Salame, Amos Tanay, Alexander van Oudenaarden, and Ido Amit. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*, 167(7):1883–1896.e15, 2016.

[72] Jonathan Jay, Jacob Bor, Elaine O. Nsoesie, Sarah K. Lipson, David K. Jones, Sandro Galea, and Julia Raifman. Neighbourhood income and physical distancing during the COVID-19 pandemic in the United States. *Nature Human Behaviour*, 4(12):1294–1302, 2020.

[73] Anna Josephson, Talip Kilic, and Jeffrey D. Michler. Socioeconomic impacts of COVID-19 in low-income countries. *Nature Human Behaviour*, 2021.

[74] Rafi Kabarriti, N. Patrik Brodin, Maxim I. Maron, Chandan Guha, Shalom Kalnicki, Madhur K. Garg, and Andrew D. Racine. Association of race and ethnicity with comorbidities and survival among patients with COVID-19 at an urban medical center in New York. *JAMA Network Open*, 3(12):e2019795, 2020.

[75] Satyen Kale. Multiarmed bandits with limited expert advice. In *Conference on Learning Theory*, pages 107–122, 2014.

[76] Kenji Kamimoto, Christy M. Hoffmann, and Samantha A. Morris. CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv*, 2020.

[77] S. Jalil Kazemitabar, Arash A. Amini, Adam Bloniarz, and Ameet S. Talwalkar. Variable importance using decision trees. In I Guyon, U V Luxburg, S Bengio, S Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 425–434. Curran Associates, Inc., 2017.

[78] Rohan Khazanchi, Evan R. Beiter, Suhas Gondi, Adam L. Beckman, Alyssa Bilinski, and Ishani Ganguli. County-level association of social vulnerability with COVID-19 cases and deaths in the USA. *Journal of General Internal Medicine*, 35(9):2784–2787, 2020.

[79] J. Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506, 1953.

[80] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *ACM Symposium on Theory of Computing*, pages 681–690, 2008.

[81] Nancy Krieger, Pamela D. Waterman, and Jarvis T. Chen. COVID-19 and overall mortality inequities in the surge in death rates by zip code characteristics: Massachusetts, January 1 to May 19, 2020. *American Journal of Public Health*, 110(12):1850–1852, 2020.

[82] Markus Kuderer, Shilpa Gulati, and Wolfram Burgard. Learning driving styles for autonomous vehicles from demonstration. In *IEEE International Conference on Robotics and Automation*, pages 2641–2646, 2015.

[83] Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 420–429, Arlington, Virginia, USA, 2014. AUAI Press.

[84] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 535–543, San Diego, California, USA, 9–12 May 2015. PMLR.

[85] T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math*, 6(1):4âĂŞ22, 1985.

[86] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[87] Joseph A. Lewnard, Vincent X. Liu, Michael L. Jackson, Mark A. Schmidt, Britta L. Jewell, Jean P. Flores, Chris Jentz, Graham R. Northrup, Ayesha Mahmud, Arthur L. Reingold, Maya Petersen, Nicholas P. Jewell, Scott Young, and Jim Bellows. Incidence, clinical outcomes, and transmission dynamics of severe coronavirus disease 2019 in California and Washington: prospective cohort study. *BMJ*, 369:m1923, 2020.

[88] Hedong Li and Gong Chen. In vivo reprogramming for CNS repair: Regenerating neurons from endogenous glial cells. *Neuron*, 91(4):728–738, 2016.

[89] Wil Lieberman-Cribbin, Stephanie Tuminello, Raja M. Flores, and Emanuela Taioli. Disparities in COVID-19 testing and positivity in New York City. *American Journal of Preventive Medicine*, 59(3):326–332, 2020.

[90] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L. Ibarra, F. Alexander Wolf, Nafissa Yakubova, Fabian J. Theis, and David Lopez-Paz. Learning interpretable cellular responses to complex perturbations in high-throughput screens. *bioRxiv*, 2021.

[91] Anna Macintyre, Daniel Ferris, Briana Gonçalves, and Neil Quinn. What has economics got to do with it? The impact of socioeconomic factors on mental

health and the case for collective action. *Palgrave Communications*, 4(1):1–5, 2018.

[92] Zachary J. Madewell, Yang Yang, Ira M. Longini, M. Elizabeth Halloran, and Natalie E. Dean. Household transmission of SARS-CoV-2: A systematic review and meta-analysis. *JAMA Network Open*, 3(12):e2031756, 2020.

[93] Nikolai Matni, Alexandre Proutiere, Anders Rantzer, and Stephen Tu. From self-tuning regulators to reinforcement learning and back again. In *IEEE Conference on Decision and Control*, pages 3724–3740, 2019.

[94] Ellicott C. Matthay, Kate A. Duchowny, Alicia R. Riley, and Sandro Galea. Projected all-cause deaths attributable to COVID-19-related unemployment in the United States. *American Journal of Public Health*, 111(4):696–699, 2021.

[95] H. Brendan McMahan and Matthew Streeter. Tighter bounds for multi-armed bandits with expert advice. In *Conference on Learning Theory*, 2009.

[96] Nadav Merlis and Shie Mannor. Tight lower bounds for combinatorial multi-armed bandits. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2830–2857. PMLR, 9–12 Jul 2020.

[97] Jerome Mertens, Maria C. Marchetto, Cedric Bardy, and Fred H. Gage. Evaluating cell reprogramming, differentiation and conversion technologies in neuroscience. *Nature Reviews Neuroscience*, 17(7):424–437, 2016.

[98] Seyed M. Moghadas, Meagan C. Fitzpatrick, Pratha Sah, Abhishek Pandey, Affan Shoukat, Burton H. Singer, and Alison P. Galvani. The implications of silent transmission for the control of COVID-19 outbreaks. *Proceedings of the National Academy of Sciences*, 117(30):17513–17515, 2020.

[99] Samantha A. Morris, Patrick Cahan, Hu Li, Anna M. Zhao, Adrianna K. San Roman, Ramesh A. Shivdasani, James J. Collins, and George Q. Daley. Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell*, 158(4):889–902, 2014.

[100] Shamim Nemati, Mohammad M. Ghassemi, and Gari D. Clifford. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2978–2981, 2016.

[101] Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 3168–3176, 2015.

[102] New York City Department of Health and Mental Hygiene. COVID-19: Data by ZIP. Data file, 2020.

[103] New York University, School of Law, Furman Center for Real Estate and Urban Policy. Eviction filings by ZIP Code. Data file, 2020.

[104] Chengzhuo Ni and Mengdi Wang. Maximum likelihood tensor decomposition of Markov decision process. In *IEEE International Symposium on Information Theory*, pages 3062–3066, 2019.

[105] Pierre Nouvellet, Sangeeta Bhatia, Anne Cori, Kylie E. C. Ainslie, Marc Baguelin, Samir Bhatt, Adhiratha Boonyasiri, Nicholas F. Brazeau, Lorenzo Cattarino, Laura V. Cooper, Helen Coupland, Zulma M. Cucunuba, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra A. Djaafara, Ilaria Dorigatti, Oliver D. Eales, Sabine L. van Elsland, Fabricia F. Nascimento, Richard G. FitzJohn, Katy A. M. Gaythorpe, Lily Geidelberg, William D. Green, Arran Hamlet, Katharina Hauck, Wes Hinsley, Natsuko Imai, Benjamin Jeffrey, Edward Knock, Daniel J. Laydon, John A. Lees, Tara Mangal, Thomas A. Mellan, Gemma Nedjati-Gilani, Kris V. Parag, Margarita Pons-Salort, Manon Ragonnet-Cronin, Steven Riley, H. Juliette T. Unwin, Robert Verity, Michaela A. C. Vollmer, Erik Volz, Patrick G. T. Walker, Caroline E. Walters, Haowei Wang, Oliver J. Watson, Charles Whittaker, Lilith K. Whittles, Xiaoyue Xi, Neil M. Ferguson, and Christl A. Donnelly. Reduction in mobility and COVID-19 transmission. *Nature Communications*, 12(1):1090, 2021.

[106] Alejandro Ocampo, Pradeep Reddy, Paloma Martinez-Redondo, Aida Platero-Luengo, Fumiyuki Hatanaka, Tomoaki Hishida, Mo Li, David Lam, Masakazu Kurita, Ergin Beyret, Toshikazu Araoka, Eric Vazquez-Ferrer, David Donoso, Jose Luis Roman, Jinna Xu, Concepcion Rodriguez Esteban, Gabriel NuÃśez, Estrella NuÃśez Delicado, Josep M. Campistol, Isabel Guillen, Pedro Guillen, and Juan Carlos Izpisua Belmonte. In vivo amelioration of age-associated hallmarks by partial reprogramming. *Cell*, 167(7):1719–1733.e12, 2016.

[107] Yutaka Okuno, Hiromi Iwasaki, Claudia S. Huettner, Hanna S. Radomska, David A. Gonzalez, Daniel G. Tenen, and Koichi Akashi. Differential regulation of the human and murine CD34 genes in hematopoietic stem cells. *Proceedings of the National Academy of Sciences*, 99(9):6246–6251, 2002.

[108] Ulrich Pfisterer, Agnete Kirkeby, Olof Torper, James Wood, Jenny Nelander, Audrey Dufour, Anders Björklund, Olle Lindvall, Johan Jakobsson, and Malin Parmar. Direct conversion of human fibroblasts to dopaminergic neurons. *Proceedings of the National Academy of Sciences*, 108(25):10343–10348, 2011.

[109] Eboni G. Price-Haywood, Jeffrey Burton, Daniel Fort, and Leonardo Seoane. Hospitalization and mortality among black patients and white patients with Covid-19. *New England Journal of Medicine*, 382:2534–2543, 2020.

[110] Li Qian, Yu Huang, C. Ian Spencer, Amy Foley, Vasanth Vedantham, Lei Liu, Simon J. Conway, Ji-dong Fu, and Deepak Srivastava. In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature*, 485(7400):593–598, 2012.

[111] Owen J. L. Rackham, Jaber Firas, Hai Fang, Matt E. Oates, Melissa L. Holmes, Anja S. Knaupp, Harukazu Suzuki, Christian M. Nefzger, Carsten O. Daub, Jay W. Shin, Enrico Petretto, Alistair R. R. Forrest, Yoshihide Hayashizaki, Jose M. Polo, Julian Gough, and The FANTOM Consortium. A predictive computational framework for direct reprogramming between human cell types. *Nature Genetics*, 48(3):331–335, 2016.

[112] Adityanarayanan Radhakrishnan, George Stefanakis, Mikhail Belkin, and Caroline Uhler. Simple, fast, and flexible framework for matrix completion with infinite width neural networks. *Proceedings of the National Academy of Sciences*, 119(16):e2115064119, 2022.

[113] Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022.

[114] Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 1952.

[115] David J. Roelfs, Eran Shor, Aharon Blank, and Joseph E. Schwartz. Misery loves company? A meta-regression examining aggregate unemployment rates and the unemployment-mortality association. *Annals of Epidemiology*, 25(5):312–322, 2015.

[116] Scott Ronquist, Geoff Patterson, Lindsey A. Muir, Stephen Lindsly, Haiming Chen, Markus Brown, Max S. Wicha, Anthony Bloch, Roger Brockett, and Indika Rajapakse. Algorithm for cellular reprogramming. *Proceedings of the National Academy of Sciences*, 114(45):11832–11837, 2017.

[117] Christopher J. Ruhm. Recessions, healthy no more? *Journal of Health Economics*, 42:17–28, 2015.

[118] Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[119] Stephanie Schmitt-Grohé, Ken Teoh, and Martín Uribe. Covid-19: Testing inequality in New York City. *Covid Economics, Vetted and Real-Time Papers*, 8:27–43, 2020.

[120] Yevgeny Seldin, Koby Crammer, and Peter Bartlett. Open problem: Adversarial multiarmed bandits with limited advice. In *Conference on Learning Theory*, pages 1067–1072, 2013.

[121] Petrônio C. L. Silva, Paulo V. C. Batista, Hélder S. Lima, Marcos A. Alves, Frederico G. Guimarães, and Rodrigo C. P. Silva. COVID-ABS: An agent-based

model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons and Fractals*, 139:110088, 2020.

[122] George A. Soper. The lessons of the pandemic. *Science*, 49(1274):501–506, 1919.

[123] Deepak Srivastava and Natalie DeWitt. In vivo cellular reprogramming: The next generation. *Cell*, 166(6):1386–1396, 2016.

[124] Kazutoshi Takahashi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–872, 2007.

[125] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, 2006.

[126] Takanao Tanaka and Shohei Okamoto. Increase in suicide following an initial decline during the COVID-19 pandemic in Japan. *Nature Human Behaviour*, 5(2):229–238, 2021.

[127] José A. Tapia Granados, James S. House, Edward L. Ionides, Sarah Burgard, and Robert S. Schoeni. Individual joblessness, contextual unemployment, and mortality risk. *American Journal of Epidemiology*, 180(3):280–287, 2014.

[128] The Lancet. Redefining vulnerability in the era of COVID-19. *The Lancet*, 395(10230):1089, 2020.

[129] The United States Centers for Disease Control and Prevention and the Office of the Assistant Secretary for Preparedness and Response. COVID-19 pandemic planning scenarios. Technical Report May 20, 2020. Available at: `https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios-archive/planning-scenarios-2020-05-20.pdf` [Accessed June 4, 2020].

[130] Matthew J. Townsend, Theodore K. Kyle, and Fatima Cody Stanford. Outcomes of COVID-19: Disparities in obesity and by ethnicity/race. *International Journal of Obesity*, 44(9):1807–1809, 2020.

[131] United States Census Bureau. Percent urban and rural in 2010 by state and county. Data file, 2012.

[132] United States Census Bureau. 2013–2017 CHAS data. Data file, 2020.

[133] United States Census Bureau. Annual estimates of the resident population by single year of age and sex for the United States: April 1, 2010 to July 1, 2019 (NC-EST2019-AGESEX-RES). Data file, 2020.

[134] United States Census Bureau. Comparative economic characteristics, 2019 American Community Survey 1-year estimates. Data file, 2020.

[135] United States Census Bureau. Selected economic characteristics, 2015–2019 American Community Survey 5-year estimates. Data file, 2020.

[136] C. A. van Diepen, J. Wolf, H. van Keulen, and C. Rappoldt. WOFOST: A simulation model of crop production. *Soil Use and Management*, 5(1):16–24, 1989.

[137] Aaron van Dorn, Rebecca E. Cooney, and Miriam L. Sabin. COVID-19 exacerbating inequalities in the US. *Lancet*, 395(10232):1243–1244, 2020.

[138] Puck B. van Kasteren, Bas van der Veer, Sharon van den Brink, Lisa Wijsman, Jørgen de Jonge, Annemarie van den Brandt, Richard Molenkamp, Chantal B. E. M. Reusken, and Adam Meijer. Comparison of seven commercial RT-PCR diagnostic kits for COVID-19. *Journal of Clinical Virology*, 128:104412, 2020.

[139] Tyler J. VanderWeele. Challenges estimating total lives lost in COVID-19 decisions: Consideration of mortality related to unemployment, social isolation, and depression. *JAMA*, 324(5):445–446, 2020.

[140] Robert Verity, Lucy C. Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick G. T. Walker, Han Fu, Amy Dighe, Jamie T. Griffin, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Anne Cori, Zulma Cucunubá, Rich FitzJohn, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Daniel Laydon, Gemma Nedjati-Gilani, Steven Riley, Sabine van Elsland, Erik Volz, Haowei Wang, Yuanrong Wang, Xiaoyue Xi, Christl A. Donnelly, Azra C. Ghani, and Neil M. Ferguson. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*, 20(6):669–677, 2020.

[141] Thomas Vierbuchen, Austin Ostermeier, Zhiping P. Pang, Yuko Kokubu, Thomas C. Südhof, and Marius Wernig. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature*, 463(7284):1035–1041, 2010.

[142] Andrew Wagenmaker, Julian Katz-Samuels, and Kevin Jamieson. Experimental design for regret minimization in linear bandits. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3088–3096. PMLR, 13–15 Apr 2021.

[143] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[144] Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5114–5122. PMLR, 10–15 Jul 2018.

[145] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[146] Joakim A. Weill, Matthieu Stigler, Olivier Deschenes, and Michael R. Springborn. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proceedings of the National Academy of Sciences*, 117(33):19658–19660, 2020.

[147] H. Weintraub, S. J. Tapscott, R. L. Davis, M. J. Thayer, M. A. Adam, A. B. Lassar, and A. D. Miller. Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proceedings of the National Academy of Sciences*, 86(14):5434–5438, 1989.

[148] Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1113–1122, Lille, France, 7–9 Jul 2015. PMLR.

[149] Bryan Wilder, Marie Charpignon, Jackson A. Killian, Han-Ching Ou, Aditya Mate, Shahin Jabbari, Andrew Perrault, Angel N. Desai, Milind Tambe, and Maimuna S. Majumder. Modeling between-population variation in COVID-19 dynamics in Hubei, Lombardy, and New York City. *Proceedings of the National Academy of Sciences*, 117(41):25904–25910, 2020.

[150] Shinya Yamanaka and Helen M. Blau. Nuclear reprogramming to a pluripotent state by three approaches. *Nature*, 465(7299):704–712, 2010.

[151] Jia Yuan Yu and Shie Mannor. Unimodal bandits. In *International Conference on International Conference on Machine Learning*, pages 41–48, 2011.

[152] Raymond Zhang and Richard Combes. On the suboptimality of Thompson sampling in high dimensions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8345–8354. Curran Associates, Inc., 2021.