Optimal Hospital Care Scheduling During the SARS-CoV-2 Pandemic

Josh C. D'Aeth[*1], Shubhechyya Ghosal[*2], Fiona Grimm[*3], David Haw[*1], Esma Koca[*2], Krystal Lau[*4], Huikang Liu[*2], Stefano Moret[*2], Dheeya Rizmie[*4], Peter C Smith[4], Giovanni Forchini[†1,5], Marisa Miraldo[†4], and Wolfram Wiesemann[†‡2]

[1]MRC Centre for Global Infectious Disease Analysis & WHO Collaborating Centre for Infectious Disease Modelling, Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), School of Public Health, Imperial College London, London, UK.

[2]Department of Analytics, Marketing & Operations, Imperial College Business School, Imperial College London, London, UK.

[3]The Health Foundation London, UK.

[4]Department of Economics and Public Policy & Centre for Health Economics and Policy Innovation, Imperial College Business School, Imperial College London, London, UK.

[5]Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden

Abstract

The COVID-19 pandemic has seen dramatic demand surges for hospital care that have placed a severe strain on health systems worldwide. As a result, policy makers are faced with the challenge of managing scarce hospital capacity so as to reduce the backlog of non-COVID patients whilst maintaining the ability to respond to any potential future increases in demand for COVID care. In this paper, we propose a nation-wide prioritization scheme that models each individual patient as a dynamic program whose states encode the patient's health and treatment condition, whose actions describe the available treatment options, whose transition probabilities characterize the stochastic evolution of the patient's health and whose rewards encode the

[*]Contributed Equally
[†]Contributed Equally
[‡]Corresponding author: ww@imperial.ac.uk

contribution to the overall objectives of the health system. The individual patients' dynamic programs are coupled through constraints on the available resources, such as hospital beds, doctors and nurses. We show that near-optimal solutions to the emerging weakly coupled counting dynamic program can be found through a fluid approximation that gives rise to a linear program whose size grows gracefully in the problem dimensions. Our case study for the National Health Service in England shows how years of life can be gained and costs reduced by prioritizing specific disease types over COVID patients, such as injury & poisoning, diseases of the respiratory system, diseases of the circulatory system, diseases of the digestive system and cancer.

Keywords: COVID, Care Prioritization, Weakly Coupled Counting Dynamic Programs, Fluid Approximation.

1 Introduction

Across health systems globally, hospitals struggle to meet the demand surges caused by the SARS-CoV-2 (hereafter COVID) pandemic. Despite the expansion of hospital capacity (*e.g.* through field hospitals; McCabe et al. 2020 and Christen et al. 2021) and the implementation of lockdown measures to smooth over time the pressures on care provision, policy makers face an unprecedented challenge in managing scarce hospital capacity and treating non-COVID patients whilst maintaining the ability to respond to any potential future increases in demand for COVID care.

In this context, care prioritization policies become vital to mitigate the morbidity and mortality associated with surges of demand overflowing the existing capacity. Prioritization policies are common in health systems where the available resources are insufficient to cope with significant seasonal demand peaks (Rizmie et al., 2019). While those pressures are normally short-lived and do not require a drastic change in care prioritization or investments in extra capacity, the pressures of the COVID pandemic are more severe due to its prolonged duration, the uncertainty in the number of COVID patients that require care, the timing of the demand surges, the intensity of resource usage required to address the needs of COVID patients and the fact that the pandemic impacts the entire population. In response, several countries have deployed a wide range of prioritization policies to delay access to care for some patients who are perceived as requiring less urgent treatment (NHS, 2020d; NICE, 2020). The National Health Service (NHS) in England, for example, provided national level guidance on the cancellation of non-urgent elective (*i.e.*, planned) procedures as well

as a prioritization to intensive care of COVID patients below the age of 65 and with a high capacity to benefit (Gardner et al., 2020). The ethical guidelines published by the German Interdisciplinary Association for Intensive Care and Emergency Medicine discuss the prioritization to intensive care of COVID patients who do not suffer from severe respiratory illness (DIVI, 2020). The Italian College of Anesthesia, Analgesia, Resuscitation and Intensive Care advised the prioritization to intensive care of COVID patients above 70 years of age that do not have more than one admission per year for a range of diseases (Riccioni et al., 2020). We refer the reader to Joebges and Biller-Andorno (2020) for a review of the prioritization guidelines applied in different countries during 2020.

The aforementioned blanket policies tend to prioritize COVID patients in detriment to patients with other diseases without systematically accounting for the trade-offs between the provision of COVID and non-COVID care. For example, non-prioritized patients that see their planned care postponed or canceled might have a higher capacity to benefit from treatment than those prioritized; also, these patients' diseases might progress considerably while they wait for care, and they may subsequently require emergency or more complex treatment, thus creating further pressures on hospital capacity. As a result, blanket policies are likely to impact morbidity and mortality as well as increase the financial burden on health systems. Against this backdrop, the Nuffield Council on Bioethics, the UK main health and healthcare ethics authority, has recently urged policy makers to develop optimal tools and national guidance to best allocate scarce hospital capacity to minimize the detrimental impact the pandemic has on population health (The Nuffield Council on Bioethics, 2020).

In this paper, we develop an optimization-based prioritization scheme that schedules patients into general & acute (G&A) as well as critical care (CC) so as to minimize overall years of life lost (YLL),[1] hospital costs or a combination of both objectives. We consider a national-level scale (rather than an individual hospital) in order to inform strategic public health policy-making, which is particularly relevant in the case of a pandemic affecting an entire country. Our optimization scheme is dynamic and considers weekly patient cohorts subdivided into different patient groups (defined by disease, age group and admission method: elective and emergency) over a 52-weeks' time horizon. We model each patient as a dynamic program (DP) whose states encode the patient's health status (proxied by the categorization elective/emergency, recovered or deceased) and treatment condition (waiting for treatment, in G&A or in CC), whose actions describe the treatment

[1] YLL quantifies the years of life lost due to premature deaths, accounting for the age at which deaths occur.

options (admit or move to G&A or to CC, deny care or discharge from hospital), whose transition probabilities characterize the stochastic evolution of the patient's health and whose rewards amount to the years of life gained, the hospital costs saved, or a combination thereof. Our model simultaneously optimizes the treatment of all patients while accounting for capacity constraints on the supply side, including the availability of G&A as well as CC beds and staff (senior doctors, junior doctors and nurses). By clustering the patients into groups (defined through the same arrival time, disease type, age group and admission type) that can each be described through the same DP, we obtain a weakly coupled counting DP that records for each patient group how many patients are in a particular state, and how many times which action is applied to those patients. We show that this weakly coupled counting DP is amenable to a fluid approximation that gives rise to a tractable linear program (LP). Moreover, the solutions to this LP allow us to recover near-optimal solutions to the weakly coupled counting DP with high probability. We demonstrate the power of our modeling framework in a case study of the NHS in England, where we cluster approximately 10 million patients (the entire population in need of care) into 3,120 patient groups whose admission we manage over the course of one year in weekly granularity.

The contributions of this paper may be summarized as follows.

(i) We develop the concept of weakly coupled counting DPs, which constitute large-scale DPs that are amenable to a tractable fluid relaxation. The fluid relaxation can be solved as an LP, and it allows recovering high-quality solutions to the weakly coupled counting DP.

(ii) We apply our findings to a case study of the NHS in England, where we show how weakly coupled counting DPs allow to prioritize access to elective and emergency care.

(iii) We publish the data of our case study as well as the source code of our prioritization scheme, so that it is available for researchers, practitioners and policy makers to develop further research and/or inform prioritization policies.

This paper is part of two related publications. The accompanying paper (D'Aeth et al., 2021) details our data collection and epidemiological modelling and uses the methodology developed in this paper to assess the effects of policies implemented by the English government to derive policy recommendations for the NHS in England. In contrast, the present paper develops the theoretical foundations of the applied analysis by D'Aeth et al. (2021): we develop the concept of weakly

coupled counting DPs, we propose their approximation via fluid DPs and the subsequent solution via LPs, and we show how to recover high-quality solutions to the original weakly coupled counting DP from solutions to these LPs.

The methodology and application of our paper builds upon a rich body of medical and methodological literature, which we review in the remainder of this section.

Under normal operation, elective care is typically scheduled via prioritization schemes (MacCormick et al., 2003; Déry et al., 2020). Recent months have seen a rapidly growing body of literature that discusses the scheduling of elective care surgeries in light of the COVID pandemic. In contrast to our work, which studies hospital care in a broader sense, the majority of that literature focuses on prioritization of COVID care (*e.g.*, Phua et al. 2020) or surgeries (mostly for cancer), with most papers evaluating the impact of the pandemic on elective surgeries (Fujita et al., 2020; Negopdiev et al., 2020; Sud et al., 2020; Yoon et al., 2020), proposing guidelines based on best practices in individual hospitals (Argenziano et al., 2020; Eichberg et al., 2020; Tzeng et al., 2020) or reviewing the guidelines of national authorities (Burki, 2020). These guidelines are developed by domain experts and tend to be qualitative in nature (Moris and Felekouras, 2020; Soltany et al., 2020). In contrast, Bertsimas, Lukin et al. (2020), Bertsimas, Pauphilet et al. (2020), Davis et al. (2020), Gao et al. (2020) and Vaid et al. (2020) employ machine learning techniques (such as support vector machines, tree ensembles and neural networks) to estimate the mortality risk of COVID patients, which can subsequently be used as a proxy of need for patient prioritization. While these contributions are important, they highlight prioritization schemes within a specific disease or sub-group of patients or care settings within a single hospital, and they are static and thus consider neither the dynamic nature of surges in demand nor the complexity of the dynamic needs of patients. In contrast, we propose a national prioritization scheme across all disease groups that accounts for future demand surges and capacity fluctuations as well as the evolution of the patients' needs over the course of the pandemic, which to the best of our best knowledge has not been proposed so far.

As an alternative to static prioritization schemes, the operations research literature has studied the dynamic management of G&A and CC capacity via admissions and discharge policies. For example, Bekker and Koeleman (2011) and Meng et al. (2015) propose capacity management policies for G&A beds using queueing theory and robust optimization, Chan et al. (2012), Kim et al. (2015)

and Ouyang et al. (2020) develop capacity management policies for CC beds via queueing theory, DPs and simulation, and Helm et al. (2011) and Shi et al. (2019) study hospital-wide capacity management policies using DPs. These approaches aim to optimize the often conflicting goals of short-term and long-term patient welfare as well as hospital costs, subject to constraints on the available resources. In contrast to our work, these papers focus on individual hospitals, which allows them to model hospital operations and within-hospital patients' care pathways at a finer granularity: admissions and discharge decisions are often taken at an hourly granularity, and some models account for longer-term implications of decisions such as readmissions.

From a methodological viewpoint, our work contributes to the rich body of research on dynamic programming. Dating back to the early work of Bellman (1952), DPs have become one of the major paradigms to model, analyze and solve dynamic decision problems affected by uncertainty (Bertsekas, 1995; Puterman, 2014). Classical DPs suffer from the *curse of dimensionality* since their state and action spaces tend to grow exponentially with the problem dimension. As a result, several methodologies have been developed by different research communities to circumvent the unfavorable scalability of classical DPs through (combinations of) decomposition and approximation techniques.

Factored Markov decision processes assume that the states of a DP can be described by assignments of values to state variables that evolve and contribute to the system's rewards largely independently, and they employ dynamic Bayesian networks to compactly represent the stochastic state evolution. The resulting optimization problems, while still exponential in size, can often be approximated well through sparse value function approximations that give rise to polynomial-time solution schemes (Boutilier et al., 1995; Guestrin et al., 2001, 2003). Translated into our context, however, the state of a factored Markov decision process would have to record the health and treatment state of millions of patients, which appears to be beyond the current state of the art in that domain (which seems to scale to tens or hundreds of state variables). Moreover, factored Markov decision processes do not offer a decomposition in terms of the actions, which is crucial in our context where the policy maker has to decide upon the treatment of each of the patients.

The literature on multi-armed and restless bandit problems studies large-scale DPs where independent components are coupled through a small number of linking constraints (Gittins et al., 2011). Since bandit problems typically assume that only one or a few arms are pulled in every round, which amounts to a single or a few patients' treatment conditions being revised at any time

point, however, their fundamental modeling assumptions appear to be at odds with our intentions.

Approximate dynamic programs offer a very general methodology to control large-scale DPs by approximating the value function through a linear combination of basis functions (Bertsekas and Tsitsiklis, 1996; Powell, 2007). While this allows to drastically reduce the number of decision variables, the number of constraints remains exponential and thus necessitates the use of additional approximation schemes such as constraint sampling (De Farias and Van Roy, 2004). More importantly, approximate dynamic programs do not exploit any specific problem structure, which is an essential feature in our problem where the different patients evolve largely independently.

More recently, approximate dynamic programs have been adapted by Hawkins (2003) and Adelman and Mersereau (2008) to weakly coupled DPs, which explicitly account for the decomposability of the overall system into largely independently evolving constituent DPs that are coupled by a small number of resource constraints. Adelman and Mersereau (2008) analyze the tightness of two approximation schemes, namely a Lagrangian relaxation that dualizes the resource constraints and an LP that imposes an additively separable value function, and they propose solution schemes based on stochastic subgradient descent and column generation. In our context, the resulting problems would be very large in scale as they would contain millions of constituent DPs, and it is unlikely that approximation guarantees similar to ours could be obtained as the constituent DPs are assumed to be pairwise different and thus cannot be aggregated to a smaller number of counting DPs.

To the best of our knowledge, the work of Bertsimas and Mišić (2016) is the closest to the methodology proposed in this paper. The authors develop a fluid approximation for large-scale DPs that decompose into largely independently evolving constituent DPs. Similar to our approximation, the authors show that their approximation is a relaxation that offers an upper bound on the optimal objective value of the original problem. Their formulation, however, scales linearly in the number of constituent DPs, which in unsuitable for our problem that comprises several million patient DPs. Moreover, since their approach can model rich dependencies between the constituent DPs (as opposed to the linear resource constraints that we employ), their action space cannot describe individual actions for each DP without incurring an exponential growth in problem size. Finally, it remains unclear whether the fluid approximation of Bertsimas and Mišić (2016) can offer performance guarantees comparable to the ones developed in this paper.

While the concept of weakly coupled counting DPs was developed with the outlined healthcare

application in mind, we emphasize its applicability in other applications as well, such as B2C marketing where current and prospective customers should be assigned to marketing campaigns based on their purchase likelihood. Here, customers can be modelled as DPs whose states encode the current product portfolio as well as preferences learnt from previous campaigns, and whose actions describe the inclusion to (or exclusion from) a particular campaign.

The remainder of the paper proceeds as follows. We introduce weakly coupled counting DPs, which will serve as our model of the health system, in Section 2. Section 3 shows that weakly coupled counting DPs are amenable to a fluid approximation that allows us to obtain high-quality solutions in polynomial time. Section 4 discusses our case study of the NHS in England, and Section 5 reports on numerical results for this case study. Section 6 concludes with a discussion of possible extensions to our model. For ease of exposition, all proofs are relegated to the appendix.

Notation. For a finite set $\mathcal{X} = \{1, \ldots, X\}$, we denote by $\Delta(\mathcal{X})$ the set of all probability distributions supported on $\mathcal{X}$, that is, all functions $p : \mathcal{X} \to \mathbb{R}_+$ satisfying $\sum_{x \in \mathcal{X}} p(x) = 1$. For a logical expression $\mathcal{E}$, we let $\mathbf{1}[\mathcal{E}] = 1$ if $\mathcal{E}$ is true and $\mathbf{1}[\mathcal{E}] = 0$ otherwise.

2 Weakly Coupled Counting Dynamic Programs

Section 2.1 shows how multiple DPs, each of which competes for the same set of resources, can be aggregated to a weakly coupled DP. Subsequently, Section 2.2 introduces the concept of a counting DP, which records how many DPs of a similar structure are in a particular state at any point in time. We will later use DPs to model individual patients and counting DPs to aggregate patients to patient groups with similar characteristics (arrival time, age group and disease type), respectively. Finally, we introduce weakly coupled counting DPs, which will allow us to combine multiple patient groups to represent our model of the entire health system.

2.1 Weakly Coupled Dynamic Programs

We model individual patients, which form the basis of our healthcare model, as DPs whose states record the patient's health (elective/emergency, recovered or deceased) and treatment state (waiting for treatment, in G&A or in CC), whose actions describe the treatment options (admit or move to G&A or to CC, deny care or discharge from hospital), whose transition probabilities characterize the stochastic evolution of the patient's health and whose rewards amount to the years of life gained,

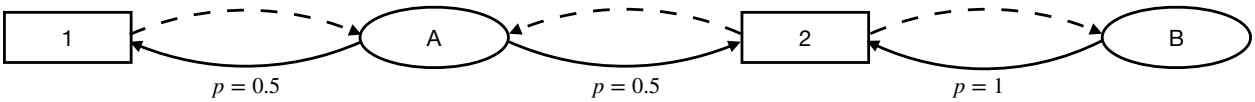


Figure 1. DP with two states and two actions. The rectangular (oval) nodes represent the states (actions). Each dashed line represents the choice of an action in a period t, whereas the solid lines represent the state transitions from period t to period $t+1$.

the costs saved, or a combination thereof.

Definition 1 (DP)**.** *For a finite time horizon $\mathcal{T} = \{1, \dots, T\}$, a* DP *is specified by the tuple $(\mathcal{S}, \mathcal{A}, q, p, r)$, where $\mathcal{S} = \{1, \dots, S\}$ denotes the finite state space, $\mathcal{A} = \{1, \dots, A\}$ is the finite action space with $\mathcal{A}_t(s) \subseteq \mathcal{A}$ the admissible actions in state $s \in \mathcal{S}$ at time $t \in \mathcal{T}$, $q \in \Delta(\mathcal{S})$ are the initial state probabilities, $p = \{p_t\}_t$ with $p_t : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, $t \in \mathcal{T}$, are the Markovian transition probabilities, and $r = \{r_t\}_t$ with $r_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $t \in \mathcal{T}$, are the expected rewards.*

In a DP, a *policy* $\pi = \{\pi_t\}_t$ with $\pi_t : \mathcal{S} \to \mathcal{A}$ specifies for each time period $t \in \mathcal{T}$ and each state $s \in \mathcal{S}$ what action $\pi_t(s) \in \mathcal{A}$ is taken. A feasible policy π must satisfy $\pi_t(s) \in \mathcal{A}_t(s)$ for all $t \in \mathcal{T}$ and $s \in \mathcal{S}$. Under the policy π, a DP evolves as follows. The initial state $\tilde{s}_1$ is random and satisfies $\mathbb{P}[\tilde{s}_1 = s] = q(s)$ for $s \in \mathcal{S}$. For $t \in \mathcal{T} \backslash \{T\}$, the transitions are governed by

$$\mathbb{P}\left[\tilde{s}_{t+1} = s'\right] \;=\; \sum_{s \in \mathcal{S}} p_t(s' \,|\, s, \pi_t(s)) \cdot \mathbb{P}\left[\tilde{s}_t = s\right] \qquad \forall s' \in \mathcal{S}.$$

The expected total reward of a policy π is $\mathbb{E}\left[\sum_{t \in \mathcal{T}} r_t(\tilde{s}_t, \pi_t(\tilde{s}_t))\right]$.

Example 1 (DP)**.** *Figure 1 illustrates a DP with the states* 1 *and* 2 *and the actions A (admissible in both states) and B (admissible in state* 2 *only). Under action A, the system transitions to either state with probability 1/2, whereas the system remains in state 2 if action B is taken. The expected rewards are $r_t(1, \mathrm{A}) = 0$ and $r_t(2, \mathrm{A}) = r_t(2, \mathrm{B}) = 1$. As a result, the unique optimal policy π takes action A in state 1 and action B in state 2, respectively.*

Our healthcare model combines all individual patient DPs to a single DP that records the state of each patient while also restricting the admissible policies to those that satisfy certain resource constraints (*e.g.*, the availability of G&A and CC beds, nurses and doctors).

Definition 2 (Weakly Coupled DP)**.** *For a finite set of DPs $(\mathcal{S}_i, \mathcal{A}_i, q_i, p_i, r_i)$, $i \in \mathcal{I} = \{1, \dots, I\}$, over the same time horizon $\mathcal{T} = \{1, \dots, T\}$, the* weakly coupled DP *$(\{\mathcal{S}_i, \mathcal{A}_i, q_i, p_i, r_i\}_i)$ is the DP*

$(\mathcal{S}, \mathcal{A}, q, p, r)$ *with state space* $\mathcal{S} = \times_{i\in\mathcal{I}} \mathcal{S}_i$*, action space* $\mathcal{A} = \times_{i\in\mathcal{I}} \mathcal{A}_i$ *with* $\mathcal{A}_t(s) = \times_{i\in\mathcal{I}} \mathcal{A}_{it}(s_i)$*,* $s \in \mathcal{S}$*, and*

$$\mathcal{A}_t^{\mathrm{C}}(s) \;=\; \left\{ a \in \mathcal{A}_t(s) \;:\; \sum_{i\in\mathcal{I}} c_{tli}(s_i, a_i) \leqslant b_{tl} \;\; \forall l \in \mathcal{L} \right\},$$

initial state probabilities $q(s) = \prod_{i\in\mathcal{I}} q_i(s_i)$ *for* $s \in \mathcal{S}$*, transition probabilities* $p_t(s' \,|\, s, a) = \prod_{i\in\mathcal{I}} p_{it}(s'_i \,|\, s_i, a_i)$ *for* $s, s' \in \mathcal{S}$ *and* $a \in \mathcal{A}$ *and expected rewards* $r_t(s, a) = \sum_{i\in\mathcal{I}} r_{it}(s_i, a_i)$*.*

Weakly coupled DPs have been studied, among others, by Hawkins (2003) and Adelman and Mersereau (2008). In a weakly coupled DP, the admissible actions $a \in \mathcal{A}_t^{\mathrm{C}}(s)$ in state $s \in \mathcal{S}$ must satisfy the constraints $a_i \in \mathcal{A}_{it}(s_i)$ of the individual DPs $i \in \mathcal{I}$ as well as the coupling resource constraints $a \in \mathcal{A}_t^{\mathrm{C}}(s)$. In particular, the feasibility of an action $a_i \in \mathcal{A}_i$ for the i-th constituent DP is not just determined by the state $s_i \in \mathcal{S}_i$, but it depends (through the resource constraints) on the states $s_{i'} \in \mathcal{S}_{i'}$, $i' \in \mathcal{I}\backslash\{i\}$, of the other constituent DPs as well. The constraints in $\mathcal{A}_t^{\mathrm{C}}$ allow us to model the resource consumption of individual patients (such as a G&A or CC bed, as well as fractions of doctor and nurse times – each of which can be modeled as a distinct resource $l \in \mathcal{L}$). Note also that the aggregation of multiple DPs to a weakly coupled DP is lossless in the sense that the state $s \in \mathcal{S}$ of a weakly coupled DP records the state $s_i \in \mathcal{S}_i$ of each constituent DP $i \in \mathcal{I}$.

In a weakly coupled DP, a *policy* $\pi = \{\pi_t\}_t$ with $\pi_t : \mathcal{S} \to \mathcal{A}$ specifies for each time period $t \in \mathcal{T}$, each DP $i \in \mathcal{I}$ and each state $s \in \mathcal{S}$ what action $[\pi_t(s)]_i \in \mathcal{A}_i$ is selected. A feasible policy π must satisfy $\pi_t(s) \in \mathcal{A}_t^{\mathrm{C}}(s)$ for all $t \in \mathcal{T}$ and $s \in \mathcal{S}$. We emphasize that the policy can choose the action $[\pi_t(s)]_i \in \mathcal{A}_i$ for the i-th DP in view of the states of all other constituent DPs, rather than just the state s_i; this is important in view of satisfying the coupling constraints. Under π, a weakly coupled DP evolves as follows. The initial state $\tilde{s}_1$ is random and satisfies $\mathbb{P}[\tilde{s}_1 = s] = \prod_{i\in\mathcal{I}} q_i(s_i) = q(s)$ for $s \in \mathcal{S}$. For $t \in \mathcal{T}\backslash\{T\}$, the transitions are governed by

$$\mathbb{P}\left[\tilde{s}_{t+1} = s'\right] \;=\; \sum_{s\in\mathcal{S}} \left[\prod_{i\in\mathcal{I}} p_{it}(s'_i \,|\, s_i, [\pi_t(s)]_i) \right] \cdot \mathbb{P}\left[\tilde{s}_t = s\right] \;=\; \sum_{s\in\mathcal{S}} p_t(s' \,|\, s, \pi_t(s)) \cdot \mathbb{P}\left[\tilde{s}_t = s\right] \qquad \forall s' \in \mathcal{S}.$$

The expected total reward of a policy π is $\mathbb{E}\left[\sum_{t\in\mathcal{T}} r_t(\tilde{s}_t, \pi_t(\tilde{s}_t))\right]$.

Example 2 (Weakly Coupled DP)**.** *Figure 3 combines the DP from Example 1 with the DP from Figure 2 to a weakly coupled DP that is subjected to the resource constraint* $\mathbf{1}[s_1 = 2 \,\wedge\, a_1 = \mathrm{B}] + \mathbf{1}[s_2 = 4 \,\wedge\, a_2 = \mathrm{D}] \leqslant 1$*, that is, at most one of the actions B and D can be selected in any*

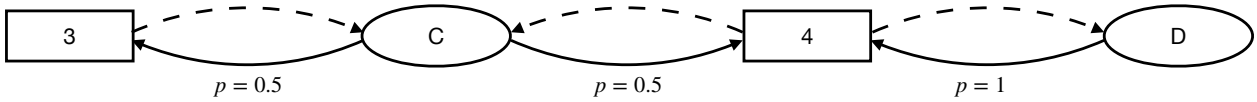


Figure 2. DP with two states 3 and 4 as well as two actions C (admissible in both states) and D (admissible in state 4 only). The expected rewards are $r_t(3, \mathrm{C}) = 0$ and $r_t(4, \mathrm{C}) = r_t(4, \mathrm{D}) = 1$.

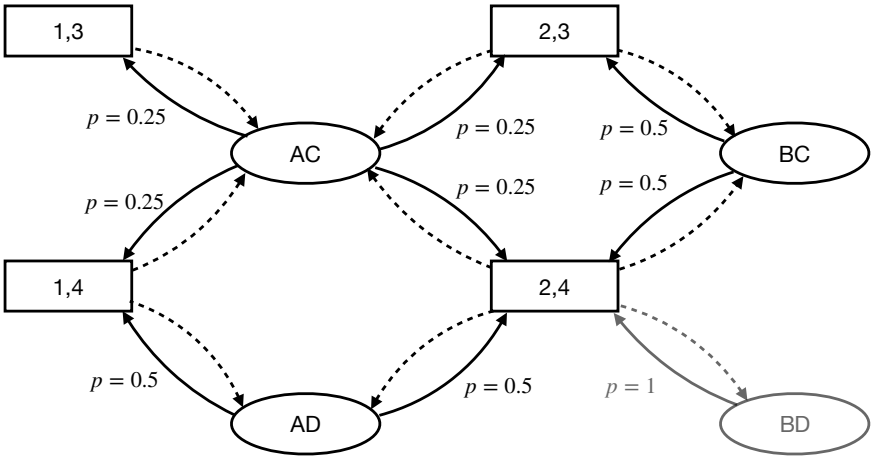


Figure 3. A weakly coupled DP is itself a DP. The values in the rectangular nodes record the states of the first and second DP, while the letters in the oval nodes denote the actions applied to each DP. The action BD violates the resource constraint.

period. As a result, any optimal policy π selects one of the actions B or D (but not both) whenever they are admissible.

2.2 Weakly Coupled Counting Dynamic Programs

Since it offers a lossless aggregation of its constituent DPs, the state and action spaces of a weakly coupled DP exhibit an undesirable scaling behavior:

$$|\mathcal{S}| \;=\; \prod_{i \in \mathcal{I}} |\mathcal{S}_i| \quad \text{and} \quad |\mathcal{A}| \;=\; \prod_{i \in \mathcal{I}} |\mathcal{A}_i|$$

The health system that we intend to model contains approximately 10 million patient DPs with 15 possible states and 6 possible actions each, thus resulting in a weakly coupled DP with approximately $15^{10{,}000{,}000}$ states and $6^{10{,}000{,}000}$ actions. Clearly, such a weakly coupled DP has to undergo a drastic dimensionality reduction before it is practical from a computational perspective.

To reduce the computational complexity of weakly coupled DPs, we first show how multiple DPs with the same state and action spaces and the same time horizon can be aggregated to a counting DP that records how many DPs are in which state at what point in time. Counting DPs will allow us to aggregate patients of the same patient group, which is characterized by the arrival time $t \in \mathcal{T}$ in our health system as well as the age group and disease type.

Definition 3 (Counting DP)**.** *For a finite time horizon $\mathcal{T} = \{1, \dots, T\}$ and n independent and identically distributed (i.i.d.) copies of a DP $(\mathcal{S}, \mathcal{A}, q, p, r)$, a* counting DP $(\mathfrak{S}, \mathfrak{A}, \mathfrak{q}, \mathfrak{p}, \mathfrak{r}; n)$ *is a DP with state space $\mathfrak{S} = \{\sigma : \mathcal{S} \to \mathbb{N}_0\}$, action space $\mathfrak{A} = \{\alpha : \mathcal{S} \times \mathcal{A} \to \mathbb{N}_0\}$, admissible actions*

$$\mathfrak{A}_t(\sigma) \;=\; \left\{ \alpha \in \mathfrak{A} \;:\; \sum_{a \in \mathcal{A}} \alpha(s,a) = \sigma(s) \;\; \forall s \in \mathcal{S}, \;\; \alpha(s,a) = 0 \;\; \forall s \in \mathcal{S}, \; \forall a \in \mathcal{A} \backslash \mathcal{A}_t(s) \right\},$$

initial state probabilities

$$\mathfrak{q}(\sigma) \;=\; \frac{n!}{\prod_{s \in \mathcal{S}} \sigma(s)!} \cdot \prod_{s \in \mathcal{S}} q(s)^{\sigma(s)} \qquad \forall \sigma \in \mathfrak{S} \;:\; \sum_{s \in \mathcal{S}} \sigma(s) = n$$

and $\mathfrak{q}(\sigma) = 0$ otherwise, transition probabilities $\mathfrak{p} = \{\mathfrak{p}_t\}_t$ with

$$\mathfrak{p}_t(\sigma' \,|\, \sigma, \alpha) \;=\; \sum_{\theta \in \Gamma(\sigma, \alpha, \sigma')} \prod_{s \in \mathcal{S}} \prod_{a \in \mathcal{A}} \left[\frac{\alpha(s,a)!}{\prod_{s' \in \mathcal{S}} \theta(s,a,s')!} \cdot \prod_{s' \in \mathcal{S}} p_t(s' \,|\, s, a)^{\theta(s,a,s')} \right]$$

with admissible transportation plans from state $\sigma \in \mathfrak{S}$ to state $\sigma' \in \mathfrak{S}$ under action $\alpha \in \mathfrak{A}$

$$\Gamma(\sigma, \alpha, \sigma') \;=\; \left\{ \theta : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{N}_0 \;:\; \sum_{s' \in \mathcal{S}} \theta(s,a,s') = \alpha(s,a) \qquad \forall s \in \mathcal{S}, \; \forall a \in \mathcal{A}, \right.$$
$$\left. \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \theta(s,a,s') = \sigma'(s') \quad \forall s' \in \mathcal{S} \right\}$$

and expected rewards $\mathfrak{r} = \{\mathfrak{r}_t\}_t$ with $\mathfrak{r}_t(\sigma, \alpha) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r_t(s,a) \cdot \alpha(s,a)$.

By construction, a counting DP can only reach states $\sigma \in \mathfrak{S}$ satisfying $\sum_{s \in \mathcal{S}} \sigma(s) = n$. Any admissible action $\alpha \in \mathfrak{A}_t(\sigma)$ thus satisfies $\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \alpha(s,a) = n$. Note that the state σ can be recovered from any admissible action α through the fact that $\sum_{a \in \mathcal{A}} \alpha(s,a) = \sigma(s)$ for all $s \in \mathcal{S}$. Thus, the transition probabilities, admissible transportation plans and expected rewards only depend on α and not on σ; to keep the notation consistent, however, we continue to include σ.

In a counting DP, a *policy* $\pi = \{\pi_t\}_t$ with $\pi_t : \mathfrak{S} \to \mathfrak{A}$ specifies for each time period $t \in \mathcal{T}$ and each counting state $\sigma \in \mathfrak{S}$ what action $\pi_t(\sigma) \in \mathfrak{A}$ is selected. A feasible policy must satisfy $\pi_t(\sigma) \in \mathfrak{A}_t(\sigma)$ for all $t \in \mathcal{T}$ and $\sigma \in \mathfrak{S}$. Under π, a counting DP evolves exactly like an ordinary DP. The initial state $\tilde{\sigma}_1$ is random and satisfies $\mathbb{P}[\tilde{\sigma}_1 = \sigma] = \mathfrak{q}(\sigma)$ for $\sigma \in \mathfrak{S}$. For $t \in \mathcal{T}\backslash\{T\}$, the transitions are governed by

$$\mathbb{P}\left[\tilde{\sigma}_{t+1} = \sigma'\right] \;=\; \sum_{\sigma \in \mathfrak{S}} \mathfrak{p}_t(\sigma' \,|\, \sigma, \pi_t(\sigma)) \cdot \mathbb{P}\left[\tilde{\sigma}_t = \sigma\right] \qquad \forall \sigma' \in \mathfrak{S}.$$

The expected total reward of a policy π is $\mathbb{E}\left[\sum_{t \in \mathcal{T}} \mathfrak{r}_t(\tilde{\sigma}_t, \pi_t(\tilde{\sigma}_t))\right]$.

To better understand Definition 3, consider n i.i.d. copies of a DP $(\mathcal{S}, \mathcal{A}, q, p, r)$ whose states and actions at time $t \in \mathcal{T}$ are recorded by the random variables $\tilde{s}_{ti}$ and $\tilde{a}_{ti}$, $i \in \mathcal{I} = \{1, \ldots, n\}$, respectively. By construction, the random quantity $|\{i \in \mathcal{I} \,:\, \tilde{s}_{1i} = s\}|$ of DPs in state $s \in \mathcal{S}$ at time period 1 follows a multinomial distribution with parameters $(n; q)$, and its probability mass function coincides with $\mathfrak{q}$ in Definition 3. In other words, the initial state $\tilde{\sigma}_1$ of the counting DP records how many of the n DPs are in each state $s \in \mathcal{S}$ at time $t = 1$. Similarly, for any fixed states $\sigma, \sigma' \in \mathfrak{S}$, action $\alpha \in \mathfrak{A}$, transportation plan $\theta \in \Gamma(\sigma, \alpha, \sigma')$ and state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the inner expression in the definition of the transition probabilities $\mathfrak{p}_t$ evaluates to

$$\begin{aligned} &\frac{\alpha(s,a)!}{\prod_{s' \in \mathcal{S}} \theta(s,a,s')!} \cdot \prod_{s' \in \mathcal{S}} p_t(s' \,|\, s, a)^{\theta(s,a,s')} \\ =\; &\mathbb{P}\Big[|\{i \in \mathcal{I} \,:\, (\tilde{s}_{ti}, \tilde{a}_{ti}, \tilde{s}_{t+1,i}) = (s,a,s')\}| = \theta(s,a,s') \;\; \forall s' \in \mathcal{S} \;\Big| \\ &\qquad\qquad |\{i \in \mathcal{I} \,:\, (\tilde{s}_{ti}, \tilde{a}_{ti}) = (s,a)\}| = \alpha(s,a)\Big], \end{aligned}$$

since, given $|\{i \in \mathcal{I} \,:\, (\tilde{s}_{ti}, \tilde{a}_{ti}) = (s,a)\}|$, the quantity $|\{i \in \mathcal{I} \,:\, (\tilde{s}_{ti}, \tilde{a}_{ti}, \tilde{s}_{t+1,i}) = (s,a,s')\}|$ follows a multinomial distribution with parameters $(\alpha(s,a); p_t(\cdot \,|\, s,a))$. Multiplying these expressions over all $(s,a) \in \mathcal{S} \times \mathcal{A}$ (since the DPs evolve independently) and summing over all transportation plans $\theta \in \Gamma(\sigma, \alpha, \sigma')$ (by the additivity of probabilities of pairwise disjoint events) shows that the transition probability $\mathfrak{p}_t(\sigma'|\sigma, \alpha)$ records the probability of the event $|\{i \in \mathcal{I} \,:\, \tilde{s}_{t+1,i} = s\}| = \sigma'(s)$, simultaneously for all $s \in \mathcal{S}$, conditional on the event $|\{i \in \mathcal{I} \,:\, (\tilde{s}_{ti}, \tilde{a}_{ti}) = (s,a)\}| = \alpha(s,a)$, simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Thus, for a given policy π the counting DP records for each time period $t \in \mathcal{T}$ how many of the n DPs are in each of the states $s \in \mathcal{S}$ under π.

Perhaps surprisingly, despite its aggregation, a feasible policy to a counting DP gives rise to feasible policies for the constituent DPs that do not incur any loss in the expected total reward.

Proposition 1. *For a DP* $(\mathcal{S}, \mathcal{A}, q, p, r)$ *and* $n \in \mathbb{N}$*, consider the corresponding counting DP* $(\mathfrak{S}, \mathfrak{A}, \mathfrak{q}, \mathfrak{p}, \mathfrak{r}; n)$ *as well as the weakly coupled DP* $(\{\mathcal{S}_i, \mathcal{A}_i, q_i, p_i, r_i\}_i)$ *with* $(\mathcal{S}_i, \mathcal{A}_i, q_i, p_i, r_i) = (\mathcal{S}, \mathcal{A}, q, p, r)$*,* $i \in \mathcal{I} = \{1, \dots, n\}$*, and no resource constraints. Fix any feasible policy* π *to the counting DP. Then any policy* π' *to the weakly coupled DP satisfying*

$$\big|\{i \in \mathcal{I} \,:\, s_i = s' \,\wedge\, \pi'_{ti}(s_i) = a'\}\big| = [\pi_t(\sigma)](s', a') \qquad \forall t \in \mathcal{T},\ \forall \sigma \in \mathfrak{S},\ \forall (s', a') \in \mathcal{S} \times \mathcal{A},$$
$$\forall s \in \mathcal{S}^n \,:\, \big[\big|\{i \in \mathcal{I} \,:\, s_i = s''\}\big| = \sigma(s'')\ \ \forall s'' \in \mathcal{S}\big]$$

is feasible, and π *and* π' *attain the same expected total reward.*

The equation in the statement of Proposition 1 ensures that for all states $s \in \mathcal{S}^n$ of the weakly coupled DP that correspond to a given state $\sigma \in \mathfrak{S}$ of the counting DP, the number of times we apply an action $a' \in \mathcal{A}$ to a DP in state $s' \in \mathcal{S}$ also coincide under π and π'.

Proposition 1 states that any policy π to a counting DP can be converted into individual policies π'_i to the constituent DPs that generate the same expected total reward. To this end, we simply need to distribute the action multiplicities $[\pi_t(\sigma)](s, a)$ for each state of the counting DP among the $\sigma(s)$ many DPs that are in state s at time t. It is noteworthy that any such distribution scheme is admissible, and they all result in the same expected total reward.

Aggregating individual patient DPs to a counting DP is *not* a lossless transformation: While a counting DP faithfully records the stochastic evolution of *the population* of individual DPs, it no longer records which state *a particular DP* is in. In other words, while the state of the i-th DP remains identifiable when the individual DPs are aggregated to a weakly coupled DP, its state is no longer identifiable by the state σ of the corresponding counting DP. In the context of our healthcare model, this implies that we have to assign the same (state/action-dependent) transition probabilities and rewards to all patients in the same patient group. This loss of flexibility is acceptable for our purpose, which is to inform a nation-wide prioritization policy (as opposed to an admissions schedule for an individual hospital), and it results in a tremendous gain in computational tractability: for $5,000$ patients of the same patient group with 15 states and 6 actions, for example, the $15^{5,000}$ states and $6^{5,000}$ actions of the associated weakly coupled DP reduce to less than $5,000^{15}$

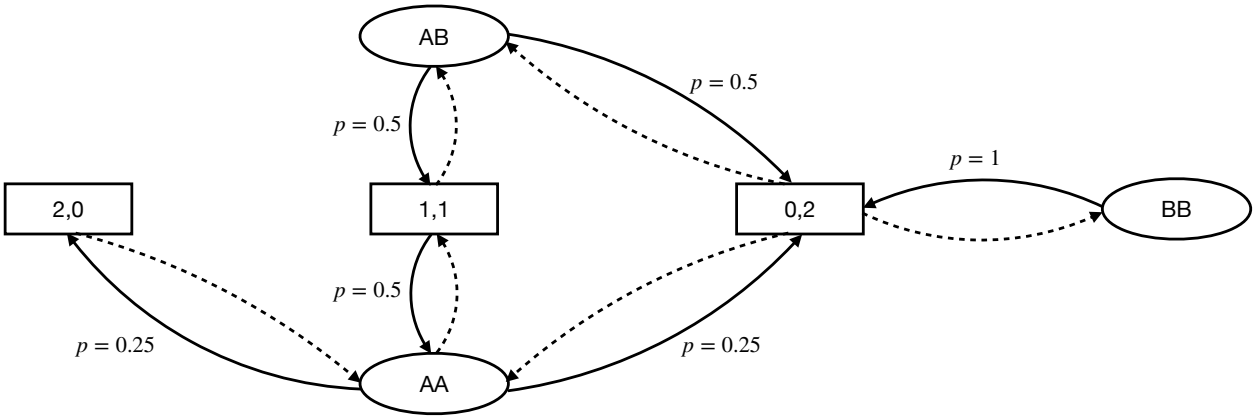


Figure 4. Counting DP for the DP from Example 1.

states and $5,000^{15 \cdot 6}$ actions for the associated counting DP.

Example 3 (Counting DP)**.** *Figure 4 presents a counting DP for $n = 2$ copies of the DP from Example 1. The states of the counting DP are characterized by two values that record the numbers of DPs in state 1 and 2, respectively, and the actions of the counting DP are characterized by the numbers of times that action A or action B is chosen.*

We now combine the concepts of weakly coupled DPs (*cf.* Definition 2) and counting DPs (*cf.* Definition 3) to weakly coupled counting DPs, which aggregate the various patient groups in our health system and restrict the admissible policies in view of the resource constraints.

Definition 4 (Weakly Coupled Counting DP)**.** *For a finite set of counting DPs $(\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j; n_j)$, $j \in \mathcal{J} = \{1, \ldots, J\}$, over the same time horizon $\mathcal{T} = \{1, \ldots, T\}$, the* weakly coupled counting DP *$(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$ is the DP $(\mathfrak{S}, \mathfrak{A}, \mathfrak{q}, \mathfrak{p}, \mathfrak{r})$ with state space $\mathfrak{S} = \times_{j \in \mathcal{J}} \mathfrak{S}_j$, action space $\mathfrak{A} = \times_{j \in \mathcal{J}} \mathfrak{A}_j$ with $\mathfrak{A}_t(\sigma) = \times_{j \in \mathcal{J}} \mathfrak{A}_{jt}(\sigma_j)$, $\sigma \in \mathfrak{S}$, and*

$$\mathfrak{A}_t^{\mathrm{C}}(\sigma) \;=\; \left\{ \alpha \in \mathfrak{A}_t(\sigma) \,:\, \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} c_{tlj}(s,a) \cdot \alpha_j(s,a) \leqslant b_{tl} \;\; \forall l \in \mathcal{L} \right\},$$

initial state probabilities $\mathfrak{q}(\sigma) = \prod_{j \in \mathcal{J}} q_j(\sigma_j)$ for $\sigma \in \mathfrak{S}$, transition probabilities $\mathfrak{p}_t(\sigma' \,|\, \sigma, \alpha) = \prod_{j \in \mathcal{J}} \mathfrak{p}_{jt}(\sigma'_j \,|\, \sigma_j, \alpha_j)$ for $\sigma, \sigma' \in \mathfrak{S}$ and $\alpha \in \mathfrak{A}$ and expected rewards $\mathfrak{r}_t(\sigma, \alpha) = \sum_{j \in \mathcal{J}} \mathfrak{r}_{jt}(\sigma_j, \alpha_j)$.

In a weakly coupled counting DP, a *policy* $\pi = \{\pi_t\}_t$ with $\pi_t : \mathfrak{S} \to \mathfrak{A}$ specifies for each time period $t \in \mathcal{T}$, each counting DP $j \in \mathcal{J}$ and each state $\sigma \in \mathfrak{S}$ what action $[\pi_t(\sigma)]_j \in \mathfrak{A}_j$

is selected. A feasible policy π must satisfy $\pi_t(\sigma) \in \mathfrak{A}_t^{\mathrm{C}}(\sigma)$ for all $t \in \mathcal{T}$ and $\sigma \in \mathfrak{S}$. Under π, a weakly coupled counting DP evolves as follows. The initial state $\tilde{\sigma}_1$ is random and satisfies $\mathbb{P}[\tilde{\sigma}_1 = \sigma] = \prod_{j \in \mathcal{J}} \mathfrak{q}_j(\sigma_j) = \mathfrak{q}(\sigma)$ for $\sigma \in \mathfrak{S}$. For $t \in \mathcal{T} \backslash \{T\}$, the transitions are governed by

$$\begin{aligned} \mathbb{P}\left[\tilde{\sigma}_{t+1} = \sigma'\right] &= \sum_{\sigma \in \mathfrak{S}} \left[\prod_{j \in \mathcal{J}} \mathfrak{p}_{jt}(\sigma'_j \,|\, \sigma_j, [\pi_t(\sigma)]_j) \right] \cdot \mathbb{P}\left[\tilde{\sigma}_t = \sigma\right] \\ &= \sum_{\sigma \in \mathfrak{S}} \mathfrak{p}_t(\sigma' \,|\, \sigma, \pi_t(\sigma)) \cdot \mathbb{P}\left[\tilde{\sigma}_t = \sigma\right] \qquad\qquad \forall \sigma' \in \mathfrak{S}. \end{aligned}$$

The expected total reward of a policy π is $\mathbb{E}\left[\sum_{t \in \mathcal{T}} \mathfrak{r}_t(\tilde{\sigma}_t, \pi_t(\tilde{\sigma}_t))\right]$.

A straightforward adaptation of Proposition 1 allows us to convert any feasible policy to the weakly coupled counting DP into policies for the constituent DPs that generate the same expected total reward. We skip the statement and proof since neither requires any new ideas.

For later reference, we define the *policy set* of a weakly coupled counting DP as $\Pi = \times_{t \in \mathcal{T}} \Pi_t$ with $\Pi_t = \{[\pi_t : \mathfrak{S} \to \mathfrak{A}] \,:\, \pi_t(\sigma) \in \mathfrak{A}_t(\sigma) \;\; \forall \sigma \in \mathfrak{S}\}$ and the *set of feasible policies* as $\Pi^{\mathrm{C}} = \times_{t \in \mathcal{T}} \Pi_t^{\mathrm{C}}$ with $\Pi_t^{\mathrm{C}} = \{\pi_t \in \Pi_t \,:\, \pi_t(\sigma) \in \mathfrak{A}_t^{\mathrm{C}}(\sigma) \;\; \forall \sigma \in \mathfrak{S}\}$, respectively. We denote the *set of state trajectories* as $\Sigma = \times_{t \in \mathcal{T}} \mathfrak{S}$; the random state evolution $\tilde{\sigma} = \{\tilde{\sigma}\}_t$ then takes values in Σ.

Example 4 (Weakly Coupled Counting DP)**.** *Figure 5 combines $n_1 = 2$ copies of the DP from Figure 1 with $n_2 = 1$ copy of the DP from Figure 2 to a weakly coupled counting DP that is subjected to the resource constraint $\alpha_1(2, \mathrm{B}) + \alpha_2(4, D) \leqslant 1$, that is, at most one of the actions B or D can be selected in any period. As a result, any optimal policy π selects either B (once) or D, but never both B and D, whenever they are admissible.*

3 Fluid Approximation

Our healthcare model will comprise approximately 10 million patients spread among $3,120$ patient groups with 15 states and 6 actions each. Assuming, for the sake of the argument, an even spread of around $3,200$ patients per patient group, the resulting weakly coupled counting DP would contain about $(3,200^{1}5)^{3,120}$ states and $(3,200^{15 \cdot 6})^{3,120}$ actions, which remains intractable. However, (weakly coupled) counting DPs lend themselves to a continuous approximation which is particularly suitable when the involved numbers of DPs are large, as is the case in our application.

To facilitate an efficient solution of the weakly coupled counting DP that represents our health-

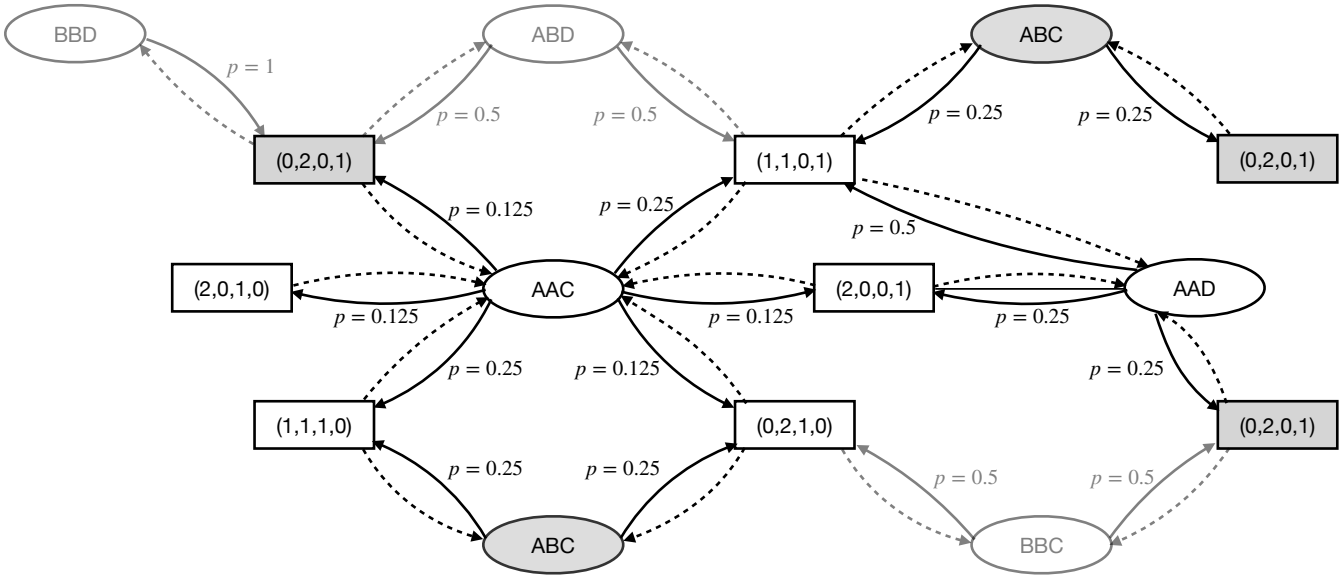


Figure 5. Weakly coupled counting DP from Example 4. The values in the rectangular nodes record how many DPs are in state $1, \ldots, 4$ (from left to right), while the letters in the oval nodes represent the actions applied to each counting DP (A and B to the first one; C and D to the second one). Grey shaded nodes are duplicated for ease of illustration. The actions ABD, BBC and BBD violate the resource constraint.

care model, Section 3.1 introduces the concept of weakly coupled k-counting DPs, which split each constituent DP into k independently evolving parts. Taking the limit as $k \longrightarrow \infty$, we arrive at the concept of a fluid limit, which treats all patients of our healthcare model as 'fluid' that flows between the different states (controlled by the policy maker's actions). Section 3.2 shows that the resulting fluid DP can be solved efficiently by an LP that scales gracefully in the dimensions of the original weakly coupled counting DP. Section 3.3, finally, shows that the LP from Section 3.2 allows us to recover high-quality solutions to the weakly coupled counting DP with high probability.

3.1 Weakly Coupled k-Counting Dynamic Programs and the Fluid Limit

We first consider a counting DP where each constituent DP is split into k parts that evolve independently of each other and that jointly account for the DP.

Definition 5 (k-Counting DP)**.** *For a finite time horizon $\mathcal{T} = \{1, \ldots, T\}$ and n i.i.d. copies of a DP $(\mathcal{S}, \mathcal{A}, q, p, r)$ formed of k i.i.d. parts each, a* k-counting DP $(\mathfrak{S}, \mathfrak{A}, \mathfrak{q}, \mathfrak{p}, \mathfrak{r}; n, k)$ *is a DP with*

state space $\mathfrak{S} = \{\sigma : \mathcal{S} \to \mathbb{N}_0/k\}$*, action space* $\mathfrak{A} = \{\alpha : \mathcal{S} \times \mathcal{A} \to \mathbb{N}_0/k\}$ *and admissible actions*

$$\mathfrak{A}_t(\sigma) \;=\; \left\{\alpha \in \mathfrak{A} \;:\; \sum_{a \in \mathcal{A}} \alpha(s,a) = \sigma(s) \;\; \forall s \in \mathcal{S}, \;\; \alpha(s,a) = 0 \;\; \forall s \in \mathcal{S}, \; \forall a \in \mathcal{A} \backslash \mathcal{A}_t(s) \right\},$$

initial state probabilities

$$\mathfrak{q}(\sigma) \;=\; \frac{[nk]!}{\prod_{s \in \mathcal{S}} [k\sigma(s)]!} \cdot \prod_{s \in \mathcal{S}} q(s)^{k\sigma(s)} \qquad \forall \sigma \in \mathfrak{S} \;:\; \sum_{s \in \mathcal{S}} \sigma(s) = n$$

and $\mathfrak{q}(\sigma) = 0$ *otherwise, transition probabilities* $\mathfrak{p} = \{\mathfrak{p}_t\}_t$ *with*

$$\mathfrak{p}_t(\sigma' \,|\, \sigma, \alpha) \;=\; \sum_{\theta \in \Gamma(\sigma,\alpha,\sigma')} \prod_{s \in \mathcal{S}} \prod_{a \in \mathcal{A}} \left[\frac{[k\alpha(s,a)]!}{\prod_{s' \in \mathcal{S}} [k\theta(s,a,s')]!} \cdot \prod_{s' \in \mathcal{S}} p_t(s' \,|\, s, a)^{k\theta(s,a,s')} \right]$$

with admissible transportation plans from state $\sigma \in \mathfrak{S}$ *to state* $\sigma' \in \mathfrak{S}$ *under action* $\alpha \in \mathfrak{A}$

$$\Gamma(\sigma, \alpha, \sigma') \;=\; \left\{ \theta : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{N}_0/k \;:\; \sum_{s' \in \mathcal{S}} \theta(s,a,s') = \alpha(s,a) \qquad \forall s \in \mathcal{S}, \; \forall a \in \mathcal{A}, \right.$$
$$\left. \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \theta(s,a,s') = \sigma'(s') \quad \forall s' \in \mathcal{S} \right\}$$

and expected rewards $\mathfrak{r} = \{\mathfrak{r}_t\}_t$ *with* $\mathfrak{r}_t(\sigma, \alpha) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r_t(s,a) \cdot \alpha(s,a)$.

Definition 5 uses the notational shorthand $\mathbb{N}_0/k = \{0, 1/k, 2/k, \ldots\}$. Note that each of the nk parts in a k-counting DP only accounts for a fraction $1/k$ of a DP, whereas each of the n parts in a counting DP accounts for an entire DP. Other than that, Definition 5 coincides with Definition 3.

A *policy* $\pi = \{\pi_t\}_t$ with $\pi_t : \mathfrak{S} \to \mathfrak{A}$ in a k-counting DP is defined analogously to a policy in an ordinary counting DP. In particular, π is feasible if $\pi_t(\sigma) \in \mathfrak{A}_t(\sigma)$ for all $t \in \mathcal{T}$ and $\sigma \in \mathfrak{S}$. Under π, the initial state $\tilde{\sigma}_1$ of the k-counting DP is random and satisfies $\mathbb{P}[\tilde{\sigma}_1 = \sigma] = \mathfrak{q}(\sigma)$ for $\sigma \in \mathfrak{S}$, and for $t \in \mathcal{T} \backslash \{T\}$ the transitions satisfy

$$\mathbb{P}\left[\tilde{\sigma}_{t+1} = \sigma'\right] \;=\; \sum_{\sigma \in \mathfrak{S}} \mathfrak{p}_t(\sigma' \,|\, \sigma, \pi_t(\sigma)) \cdot \mathbb{P}\left[\tilde{\sigma}_t = \sigma\right] \qquad \forall \sigma' \in \mathfrak{S}.$$

The expected total reward of a policy π is $\mathbb{E}\left[\sum_{t \in \mathcal{T}} \mathfrak{r}_t(\tilde{\sigma}_t, \pi_t(\tilde{\sigma}_t))\right]$.

Example 5 (k-Counting DP)**.** *For the counting DP from Example 3, Figure 6 illustrates the dis-*

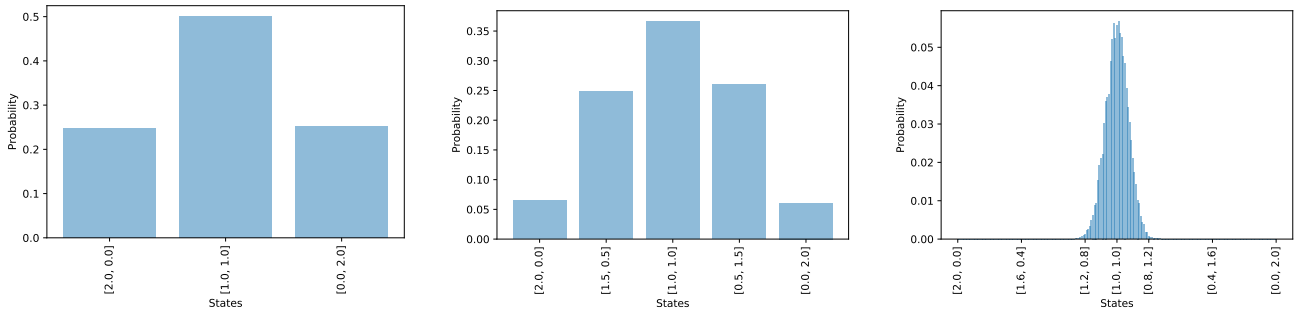


Figure 6. Next state distribution for a k-counting DP with $k = 1$ (left), $k = 2$ (middle) and $k = 100$ (right).

tribution of the next state $\tilde{\sigma}_{t+1}$ *if the current state satisfies* $\tilde{\sigma}_t = (2, 0)$ *pointwise and we implement the action* α *characterized by* $\alpha_t(s, \mathrm{A}) = \sigma_t(s)$ *and* $\alpha_t(s, \mathrm{B}) = 0$, $s \in \mathcal{S}$, *for various values of* k. *The figure shows that the next state distribution converges to a Dirac distribution that places all probability mass on the state* $\sigma_{t+1} = (1, 1)$.

Definition 6 (Weakly Coupled k-Counting DP)**.** *For a finite set of* k*-counting DPs* $(\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j; n_j, k)$, $j \in \mathcal{J} = \{1, \ldots, J\}$, *over the same time horizon* $\mathcal{T} = \{1, \ldots, T\}$, *the* weakly coupled k-counting DP $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j, k)$ *is the DP* $(\mathfrak{S}, \mathfrak{A}, \mathfrak{q}, \mathfrak{p}, \mathfrak{r})$ *with state space* $\mathfrak{S} = \times_{j \in \mathcal{J}} \mathfrak{S}_j$, *action space* $\mathfrak{A} = \times_{j \in \mathcal{J}} \mathfrak{A}_j$ *with* $\mathfrak{A}_t(\sigma) = \times_{j \in \mathcal{J}} \mathfrak{A}_{jt}(\sigma_j)$, $\sigma \in \mathfrak{S}$, *and*

$$\mathfrak{A}_t^{\mathrm{C}}(\sigma) \;=\; \left\{ \alpha \in \mathfrak{A}_t(\sigma) \;:\; \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} c_{tlj}(s, a) \cdot \alpha_j(s, a) \leqslant b_{tl} \;\; \forall l \in \mathcal{L} \right\},$$

initial state probabilities $\mathfrak{q}(\sigma) = \prod_{j \in \mathcal{J}} q_j(\sigma_j)$ *for* $\sigma \in \mathfrak{S}$, *transition probabilities* $\mathfrak{p}_t(\sigma' \,|\, \sigma, \alpha) = \prod_{j \in \mathcal{J}} \mathfrak{p}_{jt}(\sigma'_j \,|\, \sigma_j, \alpha_j)$ *for* $\sigma, \sigma' \in \mathfrak{S}$ *and* $\alpha \in \mathfrak{A}$ *and expected rewards* $\mathfrak{r}_t(\sigma, \alpha) = \sum_{j \in \mathcal{J}} \mathfrak{r}_{jt}(\sigma_j, \alpha_j)$.

In a weakly coupled k-counting DP, a *policy* $\pi = \{\pi_t\}_t$ with $\pi_t : \mathfrak{S} \to \mathfrak{A}$ specifies for each time period $t \in \mathcal{T}$, each k-counting DP $j \in \mathcal{J}$ and each state $\sigma \in \mathfrak{S}$ what action $[\pi_t(\sigma)]_j \in \mathfrak{A}_j$ is selected. A feasible policy π must satisfy $\pi_t(\sigma) \in \mathfrak{A}_t^{\mathrm{C}}(\sigma)$ for all $\sigma \in \mathfrak{S}$ and $t \in \mathcal{T}$. Under π, the initial state $\tilde{\sigma}_1$ of the weakly coupled k-counting DP is random and satisfies $\mathbb{P}[\tilde{\sigma}_1 = \sigma] = \prod_{j \in \mathcal{J}} \mathfrak{q}_j(\sigma_j) = \mathfrak{q}(\sigma)$

for $\sigma \in \mathfrak{S}$. For $t \in \mathcal{T}\backslash\{T\}$, the transitions are governed by

$$\begin{aligned}\mathbb{P}\left[\tilde{\sigma}_{t+1} = \sigma'\right] &= \sum_{\sigma\in\mathfrak{S}} \left[\prod_{j\in\mathcal{J}} \mathfrak{p}_{jt}(\sigma'_j \,|\, \sigma_j, [\pi_t(\sigma)]_j)\right] \cdot \mathbb{P}\left[\tilde{\sigma}_t = \sigma\right] \\ &= \sum_{\sigma\in\mathfrak{S}} \mathfrak{p}_t(\sigma' \,|\, \sigma, \pi_t(\sigma)) \cdot \mathbb{P}\left[\tilde{\sigma}_t = \sigma\right] \qquad\qquad \forall \sigma' \in \mathfrak{S}.\end{aligned}$$

The expected total reward of a policy π is $\mathbb{E}\left[\sum_{t\in\mathcal{T}} \mathfrak{r}_t(\tilde{\sigma}_t, \pi_t(\tilde{\sigma}_t))\right]$.

Fix a sequence of weakly coupled k-counting DPs $(\{\mathfrak{S}_j^k, \mathfrak{A}_j^k, \mathfrak{q}_j^k, \mathfrak{p}_j^k, \mathfrak{r}_j^k\}_j; \{n_j^k\}_j, k)$, $k \in \mathbb{N}$, where each constituent k-counting DP $(\mathfrak{S}_j^k, \mathfrak{A}_j^k, \mathfrak{q}_j^k, \mathfrak{p}_j^k, \mathfrak{r}_j^k; n_j^k, k)$, $j \in \mathcal{J}$, is based on the same DP $(\mathcal{S}_j, \mathcal{A}_j, q_j, p_j, r_j)$ for all $k \in \mathbb{N}$ and where $n_j^k = n_j^l$ for all $k, l \in \mathbb{N}$. Let $\tilde{s}^k_{t,(j,i,\ell)}$ be the random state of the ℓ-th part of the i-th DP in the j-th counting DP of the k-th weakly coupled k-counting DP at time t. Consistency requires that the random state evolution $\tilde{\sigma}^k = \{\tilde{\sigma}^k_t\}_t$ of the k-th weakly coupled k-counting DP satisfies $\tilde{\sigma}^k_{tj}(s) = \sum_{i=1}^{n_j} \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}[\tilde{s}^k_{t,(j,i,\ell)} = s]$ pointwise for all $k \in \mathbb{N}$, $t \in \mathcal{T}$, $j \in \mathcal{J}$ and $s \in \mathcal{S}_j$. The random initial state $\tilde{\sigma}^k_1$ of the k-th weakly coupled k-counting DP then satisfies

$$\tilde{\sigma}^k_{1j}(s) = \sum_{i=1}^{n_j} \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}[\tilde{s}^k_{1,(j,i,\ell)} = s] \underset{k\longrightarrow\infty}{\longrightarrow} \sum_{i=1}^{n_j} \mathbb{E}\big[\mathbf{1}[\tilde{s}^1_{1,(j,1,1)} = s]\big] = \sum_{i\in\mathcal{I}(j)} \mathbb{P}\big[\tilde{s}^1_{1,(j,1,1)} = s\big] = n_j \cdot q_j(s)$$

for all $j \in \mathcal{J}$ and $s \in \mathcal{S}_j$, and the convergence takes place almost surely due to the strong law of large numbers. Similarly, assume that the state of the k-th weakly coupled k-counting DP at time $t \in \mathcal{T}$ is $\tilde{\sigma}^k_t = \sigma^k_t$, that the action $\alpha^k_t \in \mathfrak{A}^k_t(\sigma^k_t)$ is taken, and that the associated states of the DP parts are $\tilde{s}^k_{t,(j,i,\ell)} = s^k_{t,(j,i,\ell)}$. Consistency then requires that $\sigma^k_{tj}(s) = \sum_{a\in\mathcal{A}_j} \alpha^k_{tj}(s,a)$ for all $k \in \mathbb{N}$, $j \in \mathcal{J}$ and $s \in \mathcal{S}_j$. Assuming that σ^k_t and α^k_t converge to σ_t and α_t as $k \longrightarrow \infty$, we observe that

$$\tilde{\sigma}^k_{t+1,j}(s') = \sum_{i=1}^{n_j} \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}[\tilde{s}^k_{t+1,(j,i,\ell)} = s'] \underset{k\longrightarrow\infty}{\longrightarrow} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s' \,|\, s, a) \cdot \alpha_{tj}(s,a)$$

for all $j \in \mathcal{J}$ and $s' \in \mathcal{S}_j$, and the convergence again takes place almost surely. Indeed, the strong law of large numbers implies that $\frac{1}{k}\sum_{\ell=1}^{k} \mathbf{1}[\tilde{s}^k_{t+1,(j,i,\ell)} = s']$ converges to its expected value, that is, the probability of $\tilde{s}^k_{t+1,(j,i,\ell)}$ being s'. The probability of $\tilde{s}^k_{t+1,(j,i,\ell)}$ being s' if $s^k_{t,(j,i,\ell)} = s$ and action $a \in \mathcal{A}_j$ is taken, on the other hand, is $p_{jt}(s' \,|\, s, a)$. As k approaches ∞, α^k_{tj} converges to α_{tj}, and thus the number of DP parts in state s that action a is applied to converges to $\alpha_{tj}(s,a) \cdot k$.

The above observation motivates the *fluid limit* that we obtain when $k \longrightarrow \infty$. In what follows,

we denote by $\delta(\cdot)$ the Dirac delta function satisfying $\delta(x) = 0$ for all $x \neq 0$ and $\int \delta(x)\,\mathrm{d}x = 1$.

Definition 7 (Fluid DP)**.** *For a weakly coupled counting DP* $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$*, the* fluid DP $(\{\overline{\mathfrak{S}}_j, \overline{\mathfrak{A}}_j, \overline{\mathfrak{q}}_j, \overline{\mathfrak{p}}_j, \overline{\mathfrak{r}}_j\}_j; \{n_j\}_j)$ *is a DP with continuous state space* $\overline{\mathfrak{S}} = \times_{j\in\mathcal{J}} \overline{\mathfrak{S}}_j$ *with* $\overline{\mathfrak{S}}_j = \{\sigma_j : \mathcal{S}_j \to \mathbb{R}_+\}$*, continuous action space* $\overline{\mathfrak{A}} = \times_{j\in\mathcal{J}} \overline{\mathfrak{A}}_j$ *with* $\overline{\mathfrak{A}}_j = \{\alpha_j : \mathcal{S}_j \times \mathcal{A}_j \to \mathbb{R}_+\}$*,* $\overline{\mathfrak{A}}_t(\sigma) = \times_{j\in\mathcal{J}} \overline{\mathfrak{A}}_{jt}(\sigma_j)$ *with*

$$\overline{\mathfrak{A}}_{jt}(\sigma_j) \;=\; \left\{ \alpha_j \in \overline{\mathfrak{A}}_j \;:\; \sum_{a\in\mathcal{A}_j} \alpha_j(s,a) = \sigma_j(s) \;\; \forall s \in \mathcal{S}_j, \;\; \alpha_j(s,a) = 0 \;\; \forall s \in \mathcal{S}_j, \; \forall a \in \mathcal{A}_j \backslash \mathcal{A}_{jt}(s) \right\}$$

and

$$\overline{\mathfrak{A}}^{\mathrm{C}}_t(\sigma) \;=\; \left\{ \alpha \in \overline{\mathfrak{A}}_t(\sigma) \;:\; \sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} c_{tlj}(s,a) \cdot \alpha_j(s,a) \leqslant b_{tl} \;\; \forall l \in \mathcal{L} \right\},$$

initial state probabilities

$$\overline{\mathfrak{q}}(\sigma) \;=\; \prod_{j\in\mathcal{J}} \prod_{s\in\mathcal{S}_j} \delta\left(\sigma_j(s) - n_j \cdot q_j(s)\right) \qquad \forall \sigma \in \overline{\mathfrak{S}}, \tag{1}$$

transition probabilities $\overline{\mathfrak{p}} = \{\overline{\mathfrak{p}}_t\}_t$ *with*

$$\overline{\mathfrak{p}}_t(\sigma' \,|\, \sigma, \alpha) \;=\; \prod_{j\in\mathcal{J}} \prod_{s'\in\mathcal{S}_j} \delta\left(\sigma_j(s') - \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_t(s'|s,a) \cdot \alpha(s,a) \right) \quad \forall \sigma, \sigma' \in \overline{\mathfrak{S}}, \; \forall \alpha \in \overline{\mathfrak{A}}, \; \forall t \in \mathcal{T} \tag{2}$$

and expected rewards $\overline{\mathfrak{r}} = \{\overline{\mathfrak{r}}_t\}_t$ *with* $\overline{\mathfrak{r}}_t(\sigma, \alpha) = \sum_{j\in\mathcal{J}} \mathfrak{r}_{jt}(\sigma_j, \alpha_j)$.

A policy $\pi = \{\pi_t\}_t$ for the fluid DP with $\pi_t : \overline{\mathfrak{S}} \to \overline{\mathfrak{A}}$ specifies for each time period $t \in \mathcal{T}$, each counting DP $j \in \mathcal{J}$ and each state $\sigma \in \overline{\mathfrak{S}}$ what action $[\pi_t(\sigma)]_j \in \overline{\mathfrak{A}}_j$ is selected. A feasible policy π must satisfy $\pi_t(\sigma) \in \overline{\mathfrak{A}}^{\mathrm{C}}_t(\sigma)$ for all $t \in \mathcal{T}$ and $\sigma \in \overline{\mathfrak{S}}$. Under π, the initial state $\tilde{\sigma}_1$ of the fluid DP satisfies $\tilde{\sigma}_{1j}(s) = n_j \cdot q_j(s)$ almost surely for all $j \in \mathcal{J}$ and $s \in \mathcal{S}_j$. For $t \in \mathcal{T}\backslash\{T\}$, under the assumption that $\tilde{\sigma}_t = \sigma_t$ almost surely, the transitions satisfy

$$\tilde{\sigma}_{t+1,j}(s') \;=\; \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_t(s'|s,a) \cdot [\pi_t(\sigma_t)]_j(s,a) \qquad \forall j \in \mathcal{J}, \; \forall s' \in \mathcal{S}_j \text{ almost surely.}$$

The expected total reward of a policy π is $\mathbb{E}\left[\sum_{t\in\mathcal{T}} \mathfrak{r}_t(\tilde{\sigma}_t, \pi_t(\tilde{\sigma}_t))\right]$.

The state and action spaces of a weakly coupled k-counting DP are of the size $\mathcal{O}\left(\prod_{j\in\mathcal{J}} (k \cdot n_j)^{|\mathcal{S}_j|}\right)$

and $\mathcal{O}\left(\prod_{j\in\mathcal{J}}(k\cdot n_j)^{|\mathcal{S}_j|\cdot|\mathcal{A}_j|}\right)$, respectively, and they thus scale exponentially in the number of involved counting DPs as well as the numbers of states and actions of the underlying DPs. In contrast, the state of the associated fluid DP can almost surely be described by a real vector of length $\sum_{j\in\mathcal{J}}|\mathcal{S}_j|$, and an action in the fluid DP is described by a real vector of length $\sum_{j\in\mathcal{J}}|\mathcal{S}_j|\cdot|\mathcal{A}_j|$. The next section will exploit this reduction in complexity to formulate the fluid DP as a linear program that scales gracefully in the parameters of the corresponding weakly coupled counting DP.

For later reference, we define the *policy set* of a fluid DP as $\overline{\Pi} = \times_{t\in\mathcal{T}}\overline{\Pi}_t$ with $\overline{\Pi}_t = \{[\pi_t : \overline{\mathfrak{S}} \to \overline{\mathfrak{A}}] \,:\, \pi_t(\sigma)\in\overline{\mathfrak{A}}_t(\sigma) \quad \forall\sigma\in\overline{\mathfrak{S}}\}$ and the *set of feasible policies* as $\overline{\Pi}^{\mathrm{C}} = \times_{t\in\mathcal{T}}\overline{\Pi}_t^{\mathrm{C}}$ with $\overline{\Pi}_t^{\mathrm{C}} = \{\pi_t\in\overline{\Pi}_t \,:\, \pi_t(\sigma)\in\overline{\mathfrak{A}}_t^{\mathrm{C}}(\sigma) \;\; \forall\sigma\in\overline{\mathfrak{S}}\}$, respectively. We denote the *set of state trajectories* as $\overline{\Sigma} = \times_{t\in\mathcal{T}}\overline{\mathfrak{S}}$; the random state evolution $\tilde{\sigma} = \{\tilde{\sigma}\}_t$ then takes values in $\overline{\Sigma}$.

3.2 Linear Programming Formulation for the Fluid Limit

We now demonstrate that an optimal policy for a fluid DP can be obtained through a linear program with $\mathcal{O}\left(|\mathcal{T}|\cdot\sum_{j\in\mathcal{J}}|\mathcal{S}_j||\mathcal{A}_j|\right)$ decision variables and $\mathcal{O}\left(|\mathcal{T}|\cdot\max\left\{\sum_{j\in\mathcal{J}}|\mathcal{S}_j|,\;|\mathcal{L}|\right\}\right)$ constraints. Our healthcare model comprises $|\mathcal{T}| = 52$ time periods (one year in weekly granularity), $|\mathcal{J}| = 3,120$ counting DPs (52 arrival times, 3 age groups and 20 disease groups), as well as $|\mathcal{S}_j| = 15$ states and $|\mathcal{A}_j| = 6$ actions per patient DP. The resulting LP, while nontrivial in size, can be solved quickly and reliably on standard hardware with off-the-shelf software.

For a weakly coupled counting DP $(\{\mathfrak{S}_j,\mathfrak{A}_j,\mathfrak{q}_j,\mathfrak{p}_j,\mathfrak{r}_j\}_j;\{n_j\}_j)$, consider the *fluid LP* defined as

$$
\begin{aligned}
\underset{\sigma,\,\pi}{\text{maximize}} \quad & \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\pi_{tj}(s,a) && \\
\text{subject to} \quad & \sigma_{1j}(s) = n_j\cdot q_j(s) && \forall j\in\mathcal{J},\ \forall s\in\mathcal{S}_j \\
& \sigma_{t+1,j}(s') = \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'\,|\,s,a)\cdot\pi_{tj}(s,a) && \forall j\in\mathcal{J},\ \forall s'\in\mathcal{S}_j,\ \forall t\in\mathcal{T}\backslash\{T\} \\
& \sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot\pi_{tj}(s,a)\leqslant b_{tl} && \forall l\in\mathcal{L},\ \forall t\in\mathcal{T} \\
& \sum_{a\in\mathcal{A}_j}\pi_{tj}(s,a) = \sigma_{tj}(s) && \forall j\in\mathcal{J},\ \forall s\in\mathcal{S}_j,\ \forall t\in\mathcal{T} \\
& \pi_{tj}(s,a) = 0 && \forall j\in\mathcal{J},\ \forall s\in\mathcal{S}_j,\ \forall a\in\mathcal{A}_j\backslash\mathcal{A}_{jt}(s),\ \forall t\in\mathcal{T} \\
& \sigma_{tj}(s),\ \pi_{tj}(s,a)\geqslant 0 && \forall j\in\mathcal{J},\ \forall(s,a)\in\mathcal{S}_j\times\mathcal{A}_j,\ \forall t\in\mathcal{T}.
\end{aligned}
\tag{3}
$$

Note that in contrast to the policy set of a fluid DP, which is infinite-dimensional, the feasible

region of the fluid LP (3) is finite-dimensional since it assigns a sequence of actions $\{\pi_t\}_t$ to a *single* state trajectory $\sigma \in \overline{\Sigma}$. This turns out to be sufficient since for a fixed policy, the fluid DP evolves according to a single state trajectory almost surely.

We first verify that the fluid LP (3) determines an optimal policy for the fluid DP, together with its associated (almost sure) state evolution.

Proposition 2. *Fix a weakly coupled counting DP* $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$.

(i) *If the fluid DP admits a feasible policy* $\overline{\pi}^0 \in \overline{\Pi}^{\mathrm{C}}$*, then a feasible solution* $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ *to the fluid LP* (3) *with objective value* θ^{LP} *gives rise to a feasible policy* $\overline{\pi} \in \overline{\Pi}^{\mathrm{C}}$ *to the fluid DP via*

$$\overline{\pi}_t(\sigma) \;=\; \begin{cases} \pi_t^{\mathrm{LP}} & \text{if } \sigma = \sigma_t^{\mathrm{LP}}, \\ \overline{\pi}_t^0(\sigma) & \text{otherwise} \end{cases} \qquad \forall t \in \mathcal{T},\ \forall \sigma \in \overline{\mathfrak{S}},$$

together with its state evolution $\tilde{\sigma} = \{\tilde{\sigma}_t\}_t$ *satisfying* $\tilde{\sigma} = \sigma^{\mathrm{LP}}$ *almost surely. Moreover, the expected total reward of* $\overline{\pi}$ *is* θ^{LP}.

(ii) *A feasible policy* $\overline{\pi} \in \overline{\Pi}^{\mathrm{C}}$ *to the fluid DP with associated state evolution* $\tilde{\sigma} = \{\tilde{\sigma}_t\}_t$ *and an expected total reward of* θ^{DP} *gives rise to a feasible solution* $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ *to the fluid LP* (3) *with objective value* θ^{DP} *via*

$$\sigma_{tj}^{\mathrm{LP}}(s') \;=\; \begin{cases} n_j \cdot q_j(s') & \text{if } t = 1, \\ \displaystyle\sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} p_{j,t-1}(s' \,|\, s, a) \cdot \pi_{t-1,j}^{\mathrm{LP}}(s, a) & \text{otherwise} \end{cases}$$

for all $t \in \mathcal{T}$, $j \in \mathcal{J}$ *and* $s' \in \mathcal{S}_j$, *as well as* $\pi_t^{\mathrm{LP}} = \overline{\pi}_t(\sigma_t^{\mathrm{LP}})$ *for all* $t \in \mathcal{T}$.

Proposition 2 immediately implies that optimal solutions to the fluid LP (3) give rise to optimal policies to the fluid DP and vice versa. The existence of a feasible policy $\overline{\pi}^0$ in part *(i)* is necessary since we require a fluid DP policy to be feasible pointwise in every state, rather than almost surely.

We now show that the fluid LP (3) constitutes a relaxation of the weakly coupled counting DP from Definition 4. While the result is intuitive, given that the associated fluid DP can be interpreted as a continuous relaxation of the weakly coupled counting DP, its proof is nontrivial since the fluid DP visits 'fractional' states that are not present in the weakly coupled counting DP.

Theorem 1. *The optimal value of the fluid LP* (3) *is greater than or equal to the optimal expected total reward of its associated weakly coupled counting DP* $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$.

Proposition 2 and Theorem 1 immediately imply that the fluid DP is a relaxation of the associated weakly coupled counting DP in the following sense.

Corollary 1. *The optimal expected total reward of a fluid DP is greater than or equal to the optimal expected total reward of its associated weakly coupled counting DP.*

3.3 Approximation Guarantees

An optimal solution $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ to the fluid LP (3) does not only give rise to an optimal policy for the fluid DP, but it also allows us to construct near-optimal policies for the weakly coupled counting DP. In this section, we study two such constructions: one that is based on a deterministic rounding scheme (Section 3.3.1) and one that is based on a randomization approach (Section 3.3.2).

3.3.1 Deterministic Rounded Policies

Fix a weakly coupled counting DP $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$ as well as an optimal solution $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}}) \in \overline{\Sigma} \times \overline{\mathfrak{A}}^T$ to its associated fluid LP (3). We consider the policy

$$\pi^\star \;=\; \Pr\left[\mathrm{L}(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})\right],$$

where the lifting operator L maps $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ to a policy $\overline{\pi} = \mathrm{L}(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ for the fluid DP via

$$[\overline{\pi}_t(\sigma)]_j(s,a) \;=\; \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{\sigma^{\mathrm{LP}}_{tj}(s)} \cdot \sigma_{tj}(s) \qquad \forall t \in \mathcal{T},\ \forall \sigma \in \overline{\mathfrak{S}},\ \forall j \in \mathcal{J},\ \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j,$$

where we adopt the convention that $0/0 = 0$, and the projection operator Pr maps $\overline{\pi}$ to a policy $\pi^\star$ for the weakly coupled counting DP according to

$$[\Pr(\overline{\pi})]_t(\sigma) \in \operatorname*{arg\,min}_{\alpha \in \mathfrak{A}_t(\sigma)} \left\|\overline{\pi}_t(\sigma) - \alpha\right\|_1 \qquad \forall t \in \mathcal{T},\ \forall \sigma \in \mathfrak{S},$$

where $\left\|\overline{\pi}_t(\sigma) - \alpha\right\|_1 \;=\; \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} \left|[\overline{\pi}_t(\sigma)]_j(s,a) \;-\; \alpha_j(s,a)\right|$. We will see below that the minimum in the above equation is indeed attained. Intuitively speaking, the lifted policy $\overline{\pi}$ employs

the actions in each counting state σ with the same frequency as π^{LP} does in the almost sure trajectory σ^{LP} from the fluid LP in the same time period, and the projected policy $\pi^\star$ rounds this policy to the nearest discrete policy that obeys the constraints of $\overline{\Pi}$. We note that neither $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ nor $\pi^\star$ is unique in general. In the remainder, all statements apply to *any* policy $\pi^\star$ emerging from the above construction.

Recall from Proposition 2 *(i)* that feasible solutions to the fluid LP (3) give rise to feasible policies to the fluid DP, provided that $\overline{\Pi}^{\mathrm{C}}$ is non-empty. Since we did not assume non-emptiness of $\overline{\Pi}^{\mathrm{C}}$ here, the lifted policy $\overline{\pi}$ may violate the resource constraints of the fluid DP.

Observation 1. *The lifted policy $\overline{\pi}$ satisfies $\overline{\pi} \in \overline{\Pi}$, but it may not be contained in $\overline{\Pi}^{\mathrm{C}}$.*

The proof of Observation 1 implies that the projected policy $\pi^\star$ may violate the resource constraints of the weakly coupled counting DP as well. We next show, however, that $\pi^\star$ is contained in the policy set Π of the weakly coupled counting DP, and that it is close to the lifted policy $\overline{\pi}$.

Proposition 3. *We have $\pi^\star \in \Pi$ as well as $\|\pi^\star_t(\sigma) - \overline{\pi}_t(\sigma)\|_\infty < 1$ for all $t \in \mathcal{T}$ and $\sigma \in \mathfrak{S}$.*

Despite potentially violating the resource constraints of Π^{C}, we now show that under suitable assumptions, the projected policy $\pi^\star$ is close to Π^{C} with high probability. The proximity of $\pi^\star$ to Π^{C} will crucially depend on the two quantities

$$\frac{\overline{p}_j^t - 1}{\overline{p}_j - 1}, \quad \text{where} \quad \overline{p}_j \;=\; \max\left\{ \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} p_{jt}(s'|s,a) \;:\; t \in \mathcal{T},\; s' \in \mathcal{S}_j \right\},\; j \in \mathcal{J},$$

as well as

$$\epsilon \;=\; \max\left\{ \sqrt{\frac{\log n_j}{2n_j}} \;:\; j \in \mathcal{J} \right\}.$$

Note that $\overline{p}_j \in [0,\, |\mathcal{S}_j| \cdot |\mathcal{A}_j|]$ by construction, and thus

$$\frac{\overline{p}_j^t - 1}{\overline{p}_j - 1} \;\in\; \left[1,\, \frac{[|\mathcal{S}_j| \cdot |\mathcal{A}_j|]^t - 1}{|\mathcal{S}_j| \cdot |\mathcal{A}_j| - 1}\right] \;\approx\; \left[1,\, [|\mathcal{S}_j| \cdot |\mathcal{A}_j|]^{t-1}\right].$$

We emphasize that $(\overline{p}_j^t - 1)\,/\,(\overline{p}_j - 1)$ does not grow with the numbers n_j of DPs in each counting DP. The quantity ϵ, on the other hand, vanishes quickly when $n_j \to \infty$ for all $j \in \mathcal{J}$.

We are now ready to analyze the performance of $\pi^\star$ in the weakly coupled counting DP.

Theorem 2 (Rounded Policy; Expected Total Reward)**.** *Denote by $\theta^\star$ and θ^{DP} the expected total reward of the rounded policy $\pi^\star$ and an optimal policy for the weakly coupled counting DP $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$, respectively. We then have*

$$\theta^\star \;\geqslant\; \theta^{\mathrm{DP}} \;-\; \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} r_{jt}(s,a) \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1}$$

as well as, with probability at least $1 - 2|\mathcal{T}| \cdot \sum_{j\in\mathcal{J}} |\mathcal{S}_j|/n_j$ for all $t \in \mathcal{T}$ and $l \in \mathcal{L}$ simultaneously,

$$\sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} c_{tlj}(s,a) \cdot [\pi_t^\star(\tilde{\sigma}_t^\star)]_j(s,a) \leqslant b_{tl} + \sum_{j\in\mathcal{J}} (1+\epsilon n_j) \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} c_{tlj}(s,a) \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1},$$

where $\tilde{\sigma}^\star = \{\tilde{\sigma}_t^\star\}_t$ is the random state evolution of the weakly coupled counting DP under $\pi^\star$.

Theorem 3 (Rounded Policy; Worst-Case Total Reward)**.** *Denote by $\tilde{\theta}^\star$ the random total reward of the rounded policy $\pi^\star$ and by θ^{DP} the expected total reward of an optimal policy for the weakly coupled counting DP $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$, respectively. We then have*

$$\tilde{\theta}^\star \;\geqslant\; \theta^{\mathrm{DP}} \;-\; \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} r_{jt}(s,a) \cdot (1+\epsilon n_j) \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1}$$

as well as, for all $t \in \mathcal{T}$ and $l \in \mathcal{L}$ simultaneously,

$$\sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} c_{tlj}(s,a) \cdot [\pi_t^\star(\tilde{\sigma}_t^\star)]_j(s,a) \leqslant b_{tl} + \sum_{j\in\mathcal{J}} (1+\epsilon n_j) \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} c_{tlj}(s,a) \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1},$$

both with probability at least $1 - 2|\mathcal{T}| \cdot \sum_{j\in\mathcal{J}} |\mathcal{S}_j|/n_j$, where $\tilde{\sigma}^\star = \{\tilde{\sigma}_t^\star\}_t$ is the random state evolution of the weakly coupled counting DP under $\pi^\star$.

To interpret the above performance guarantees in light of our healthcare model, we consider an asymptotic setting where $n_j \to \infty$ for all $j \in \mathcal{J}$, and where the available resources $b_{tl} \propto \sum_{j\in\mathcal{J}} n_j$ scale in the number of patients to be treated. One can verify that in this case, the expected total reward $\theta^{\mathrm{DP}} \propto \sum_{j\in\mathcal{J}} n_j$ of the weakly coupled counting DP under the optimal policy scales with the number of patients as well. In contrast, we assume that the number $|\mathcal{T}|$ of time periods, the number $|\mathcal{J}|$ of patient groups as well as the number of states $|\mathcal{S}_j|$ and actions $|\mathcal{A}_j|$ per patient group remain constant. In that setting, Theorem 2 shows that the expected total reward $\theta^\star$ of

the rounded policy $\pi^\star$ equals the optimal expected total reward θ^{DP} minus a constant term, since the expression subtracted on the right-hand side of the objective bound does not grow with n_j, $j \in \mathcal{J}$. In contrast, Theorems 2 and 3 show that the worst-case total reward as well as the resource consumptions deviate from the optimal expected total reward θ^{DP} and the resource availabilities b_{tl}, respectively, by constants multiplied with ϵn_j, where $\epsilon n_j \propto \sqrt{n_j \log n_j}$. While these expressions are no longer constants, they are sublinear and thus imply that the worst-case suboptimality as well as the resource violations, on a percentage basis, vanish as well when the number of patients grows. The results also reveal the price to be paid when moving from a guarantee in expectation (*cf.* Theorem 2) to a guarantee in terms of the worst case (*cf.* Theorem 3): The suboptimality increases from an additive constant to a sublinear expression.

A question of practical concern is how the policy $\pi^\star$ to the weakly coupled counting DP can be computed from a solution $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ to the fluid LP (3). This appears difficult as the lifted policy $\mathrm{L}(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ lives in an infinite-dimensional space and the projection operator Pr seems to require the solution of a combinatorial optimization problem. Fortunately, given $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$, the policy $\pi^\star$ affords a simple characterization in closed form.

Proposition 4. *For all $t \in \mathcal{T}$, $\sigma \in \mathfrak{S}$, $j \in \mathcal{J}$ and $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$, the policy $\pi^\star$ satisfies*

$$\left[\pi_t^\star(\sigma)\right]_j (s,a) \;=\; \lfloor [\overline{\pi}_t(\sigma)]_j(s,a) \rfloor + \mathbf{1}[(s,a) \in \mathcal{I}_{tj}(\sigma)],$$

where $\mathcal{I}_{tj}(\sigma) \subseteq \mathcal{S}_j \times \mathcal{A}_j$ contains the pairs $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$ corresponding to the $\|\mathrm{frac}([\overline{\pi}_t(\sigma)]_j(s,\cdot))\|_1$ largest components of $\mathrm{frac}([\overline{\pi}_t(\sigma)]_j(s,\cdot))$.

Proposition 4 denotes by $\mathrm{frac}(x) = x - \lfloor x \rfloor$ the fractional part of a number $x \in \mathbb{R}$, which we apply to functions component-wise. Note that $\|\mathrm{frac}([\overline{\pi}_t(\sigma)]_j(s,\cdot))\|_1 \in \mathbb{N}_0$ since $\sum_{a \in \mathcal{A}_j} [\overline{\pi}_t(\sigma)]_j(s,a) = \sigma_j(s) \in \mathbb{N}_0$ for all $\sigma \in \mathfrak{S}$ and all $s \in \mathcal{S}_j$. We emphasize that the set $\mathcal{I}_{tj}(\sigma)$ is not necessarily unique.

3.3.2 Randomized Policies

We now utilize an optimal solution $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ to the fluid LP (3) to construct a *randomized* policy for our health system. Our analysis will crucially rely on the actions applied to each constituent DP being independent. We therefore cannot operate on the weakly coupled counting DP, which abstracts away from the dependence structure between the constituent DPs. Instead, we transform

the weakly coupled counting DP $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$ representing our healthcare model to a weakly coupled DP $(\{\mathcal{S}_i, \mathcal{A}_i, q_i, p_i, r_i\}_i)$ as follows. We use the same time horizon $\mathcal{T}$, we set $\mathcal{I} = \bigcup_{j \in \mathcal{J}} \{(j,i) : i = 1, \ldots, n_j\}$, $\mathcal{S}_{(j,i)} = \mathcal{S}_j$, $\mathcal{A}_{(j,i)} = \mathcal{A}_j$, $\mathcal{A}_{t,(j,i)}(s) = \mathcal{A}_{tj}(s)$, $s \in \mathcal{S}_j$, $q_{(j,i)} = q_j$ and $p_{t,(j,i)} = p_{tj}$ for all $t \in \mathcal{T}$ and $(j,i) \in \mathcal{I}$. The weakly coupled DP thus records the state and action of the i-th DP in the j-th counting DP, $(j,i) \in \mathcal{I}$, in the (j,i)-th DP explicitly, whereas the weakly coupled counting DP aggregates the DPs $(j,1), \ldots, (j,n_j)$ to the j-th counting DP, $j \in \mathcal{J}$.

Recall that a deterministic policy $\pi = \{\pi_t\}_t$, $\pi_t : \mathcal{S} \to \mathcal{A}$, for the weakly coupled DP assigns an action $a \in \mathcal{A}$ to each possible state $s \in \mathcal{S}$ for each time period $t \in \mathcal{T}$. We now consider randomized policies $\pi = \{\pi_t\}_t$, $\pi_t : \mathcal{S} \to \Delta(\mathcal{A})$, for the weakly coupled DP that assign probability distributions over all actions $a \in \mathcal{A}$ to each possible state $s \in \mathcal{S}$ for each time period $t \in \mathcal{T}$.

Fix a weakly coupled counting DP $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$ as well as an optimal solution $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ to its associated fluid LP (3). We consider the following randomized policy $\pi^\star = \{\pi_t^\star\}_t$:

$$[\pi_t^\star(s)](a) \;=\; \prod_{j \in \mathcal{J}} \prod_{i=1}^{n_j} \frac{\pi_{tj}^{\mathrm{LP}}(s_{(j,i)}, a_{(j,i)})}{\sigma_{tj}^{\mathrm{LP}}(s_{(j,i)})} \qquad \forall t \in \mathcal{T},\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

Again, we adopt the convention that $0/0 = 0$. Intuitively speaking, the randomized policy $\pi^\star$ considers each constituent DP $(j,i) \in \mathcal{I}$ independently, and it employs each action $a_{(j,i)} \in \mathcal{A}_{(j,i)}$ with a probability that equals the fraction of times this action has been selected in the fluid LP (3) under the almost sure trajectory in the same time period.

We now study the performance of $\pi^\star$ in the weakly coupled DP. Our results use the random states $\tilde{s}_{t,(j,i)}$ of the (j,i)-th DP and the random actions $\tilde{a}_{t,(j,i)}$ applied to this DP in time period t.

Theorem 4 (Randomized Policy; Expected Total Reward)**.** *Denote by $\theta^\star$ and θ^{DP} the expected total reward of the randomized policy $\pi^\star$ and an optimal policy for the weakly coupled counting DP $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$, respectively. We then have*

$$\theta^\star \;\geqslant\; \theta^{\mathrm{DP}}$$

as well as, with probability at least $1 - 2|\mathcal{T}| \cdot \sum_{j \in \mathcal{J}} |\mathcal{S}_j| \cdot |\mathcal{A}_j| \,/\, n_j$,

$$\sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} c_{tlj}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)}) \leqslant b_{tl} + \epsilon \cdot \sum_{j \in \mathcal{J}} n_j \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} c_{tlj}(s,a) \qquad \forall t \in \mathcal{T},\ \forall l \in \mathcal{L}.$$

Theorem 5 (Randomized Policy; Worst-Case Total Reward)**.** *Denote by* $\tilde{\theta}^\star$ *the random total reward of the randomized policy* $\pi^\star$ *and by* θ^{DP} *the expected total reward of an optimal policy for the weakly coupled counting DP* $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$*, respectively. We then have*

$$\tilde{\theta}^\star \;\geqslant\; \theta^{\mathrm{DP}} \;-\; \epsilon \cdot \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} n_j \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} r_{jt}(s,a)$$

as well as

$$\sum_{j\in\mathcal{J}} \sum_{i=1}^{n_j} c_{tlj}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)}) \leqslant b_{tl} + \epsilon \cdot \sum_{j\in\mathcal{J}} n_j \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} c_{tlj}(s,a) \qquad \forall t \in \mathcal{T},\, \forall l \in \mathcal{L},$$

both with probability at least $1 - 2|\mathcal{T}| \cdot \sum_{j\in\mathcal{J}} |\mathcal{S}_j| \cdot |\mathcal{A}_j| \,/\, n_j$*.*

The bounds of Theorems 4 and 5 are stronger than those of Theorems 2 and 3 in the sense that the expected total reward bound does not contain any additive constants and the worst-case total reward and the resource violation bounds contain smaller additive expressions (which nevertheless exhibit the same asymptotic behavior). On the flip side, the rounded policy from the previous section offers more implementational leeway since it can be converted into many different policies for the constituent DPs that all generate the same expected total reward (*cf.* Proposition 1), whereas the randomized policy requires an i.i.d. application of the actions across the constituent DPs.

4 Elective Care Scheduling in England

In this section, we describe our case study of the NHS in England. In particular, we discuss the overall setup of our case study (Section 4.1), the employed data sources (Section 4.2) as well as the DPs that model the patients of our healthcare model (Section 4.3).

4.1 Experimental Setup

We apply our framework to the NHS in England during the COVID pandemic. We aim to optimally schedule elective procedures over a 56-week planning horizon (52 reported weeks plus an additional 4 weeks to avoid end-of-time-horizon effects), starting from March 2, 2020, with the objective of minimizing YLL. We consider a total of 10.45 million non-COVID patients that are subdivided into *(i)* electives (3.9 millions) and emergencies (6.55 millions), *(ii)* 20 disease groups and *(iii)* 3 age

groups (under 25 years, 25–64 years, over 64 years). We also consider 349,279 COVID patients, all of whom are emergencies.

Our model has a weekly granularity. At the beginning of each week, a new inflow of patients in need of elective and emergency care (hereafter denoted as elective and emergency patients, respectively) enters the system. Strictly speaking, our model distinguishes between different medical procedures, and thus one and the same patient in need of several procedures is included as multiple different patients in our model. For ease of exposition, however, we continue to talk about patients in the follwing. Patients are then admitted to hospital, and they evolve over the duration of the week. Emergencies are always admitted to hospital upon arrival (if capacity permits). Elective patients who are not immediately admitted to hospital wait in a queue. While in the queue, an elective patient's condition might worsen and hence require emergency admission. Based on the severity of their conditions and resource availability, patients are first admitted to G&A or to CC and can transition between G&A and CC in the following weeks of hospitalization until they are eventually discharged from hospital (recovered or deceased). Transitions between G&A and CC are decided upon at the beginning of each week, and they are based on transition probabilities that are specific to the different patient groups and admission types (elective or emergency).

4.2 Data Sources

We combine a large set of modeling and administrative data to create a comprehensive open-source dataset for the NHS in England that includes elective and emergency patient inflows, transition probabilities, availability and requirement of resources and cost of care. To this end, we leverage several data sources. Non-COVID patient inflows are forecasted with local linear trend models with trigonometric seasonality time series methods using individual level patient records across all hospitals in England from the Hospital Episode Statistics (HES), see NHS Digital (2020). HES contain patient level data with diagnoses, individual characteristics, care received, date of admission and time and mode of discharge from hospital.

Weekly inflows of COVID patients are generated by a susceptible-exposed-infected-recovered dynamic transmission model of SARS-CoV-2. Epidemic projections are made using the integrated epidemic/economic model Daedalus (Haw et al., 2020), in which the population consists of four age groups: pre-schoolers, school-age children, working-age adults and retired. The working-age

population is further divided into 63 economic sectors plus non-working adults. Each of these groups is further divided into eight subgroups with respect to the disease status: the susceptible, the exposed, the asymptomatic infectious, the infected with mild symptoms, the infected with influenza-like symptoms, the hospitalized, the recovered and the dead. The model fits four parameters to English hospital occupancy data (NHS, 2020b) from March 20, 2020 to June 30, 2020: epidemic onset, basic reproductive number, lockdown onset, and reduction in transmission during lockdown due to pandemic mitigation strategies. In our numerical studies, we consider an epidemiological scenario defined by a lockdown enforced on January 1, 2021 and the maximum value of the reproductive number $R_{\max} = 1.2$ attained during the post-lockdown period.

The evolution of hospitalized non-COVID and COVID patients is modeled with Kaplan-Meier estimators using, respectively, HES data and individual clinical data from patients who received care at the Imperial College Healthcare NHS Trust (Perez-Guzman et al., 2020). Staff resources are calculated from the 2020 NHS Electronic Staff Records dataset, and G&A and CC bed availabilities are obtained from the February 2020 KH03-Quarterly Bed Availability and Occupancy Dataset and from the February 2020 Critical Care Monthly Situation Reports datasets (NHS, 2020a,c). Staff-to-bed ratios are calculated using the Royal College of Physicians guidance (RCP London, 2020; Royal College of Nursing, 2020). Our model always considers properly staffed beds (accounting for the applicable staff-to-bed ratios); indeed, staff has emerged as a key resource bottleneck during the pandemic. YLL are calculated using standard life tables data provided by the Office for National Statistics (Office for National Statistics, 2019). Patients are individually costed using the National Cost Collection dataset from 2015 to 2019 (NHS England, 2020) matched to HES data at Health Resource Group level (HRGs equivalent to Diagnosis Related Groups international coding system).

Table 1 and Figure 7 summarize the main input data for our case study. Table 1 reports the availability of beds and staff in G&A and CC across the NHS in England, together with the corresponding staff-to-bed ratios. Figure 7 shows the weekly inflows of elective and emergency patients from March 2020 to February 2021, categorized by disease type according to the International Classification of Diseases (ICD) standard. The figure only displays the five largest patient groups individually, whereas the smaller remaining groups are collective referred to as "Others". We observe a seasonal trend in patient inflows as well as dips in January, March/April, May/June and September that are associated with winter/weather, school and long bank holiday effects.

Table 1. Availability of resources and staff-to-bed ratios.

	Capacity				Staff-to-bed Ratio		
	Beds	Senior Doctors	Junior Doctors	Nurses	Senior Doctors	Junior Doctors	Nurses
G&A	102,186	10,764	8,539	43,214	15	15	5
CC	4,122	1,013	963	18,856	15	8	1

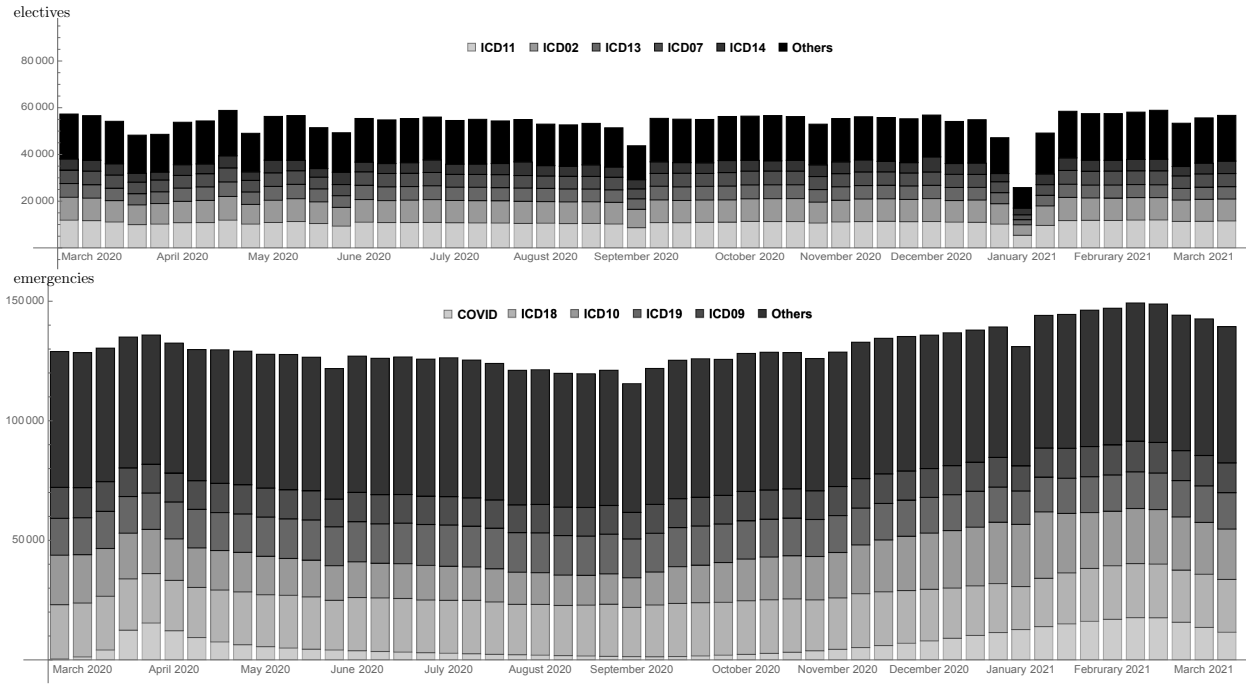


Figure 7. Weekly inflows of elective (top) and emergency (bottom) patients categorized by disease group (ICD02: neoplasms; ICD07: diseases of the eye and adnexa; ICD09: diseases of the circulatory system; ICD10: diseases of the respiratory system; ICD11: diseases of the digestive system; ICD13: diseases of the musculoskeletal system and connective tissue; ICD14: diseases of the genitourinary system; ICD18: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; ICD19: injury, poisoning and certain other consequences of external causes).

All data are made available open-source;[2] additional information on the data sources as well as the data treatment methodology is provided by D'Aeth et al. (2021).

4.3 Dynamic Programming Model of an Individual Patient

Recall that the patients in our healthcare model are partitioned into 3,120 groups, each of which is characterized by an arrival time in the system (52 weeks), one out of 20 disease types and one out of 3 age groups. We associate a DP with each of these patient groups. All DPs share the same

[2]Source code and data available at: `https://github.com/ImperialCollegeLondon/dp2lp`

state and action sets, but the DPs of different patient groups differ in their admissible actions per time period and state, their transition probabilities as well their expected rewards.

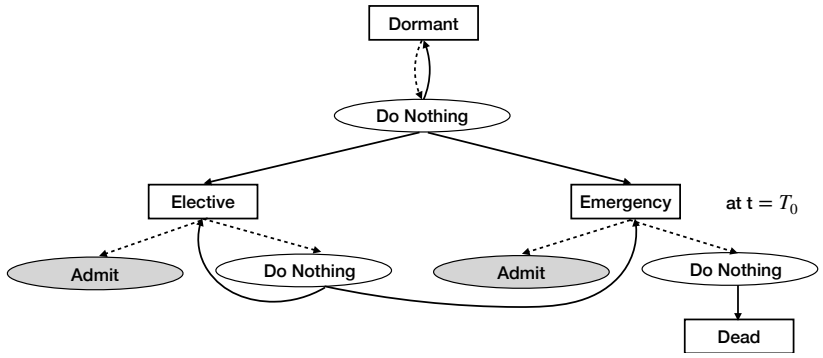
Figure 8 offers a schematic representation of a patient DP. Apart from the patients that enter our healthcare model in the first week, each patient is *Dormant* until the beginning of week $t = T_0$, at which point she enters the system either as to be admitted for a planned procedure (*Elective*) or as emergency (*Emergency*). Emergency patients are admitted to hospital immediately if capacity permits; we assume that emergency patients who are denied admission die, which is represented by the *Dead* state. Elective patients that are not immediately admitted to hospital, on the other hand, remain waiting and run a risk of requiring emergency care in subsequent weeks.

Upon admission to hospital as elective or emergency, a patient requires either G&A (*Initially requires G&A* state) or CC (*Initially requires CC* state). A patient in need of G&A is admitted to G&A (*G&A* state), whereas a patient requiring CC can be assigned to either CC (*CC* state) or, in case of capacity shortages, to G&A. In the latter case, the transition into a designated G^* state implies that the patient subsequently evolves according to a different set of transition probabilities that account for an increased mortality risk associated with the denial of CC.

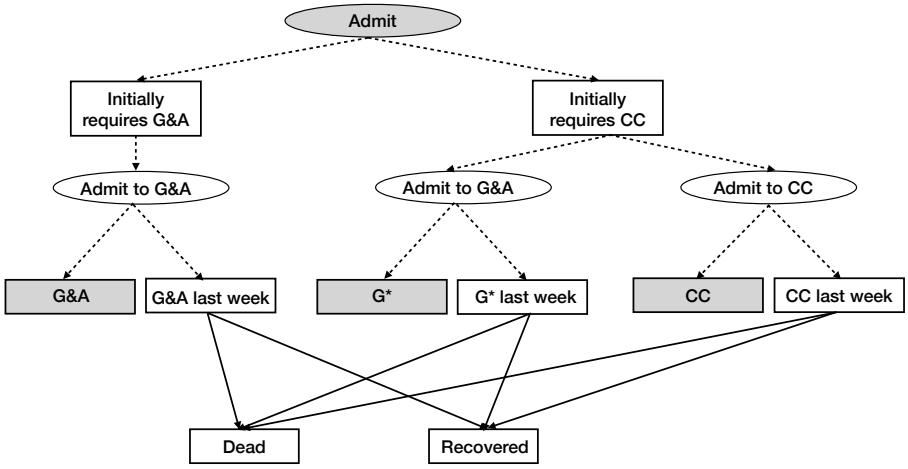
In the following weeks, depending on her response to the treatment, the patient can either require the same care regime or be moved to another one (*cf.* the *Requires G&A* and *Requires CC* states). Depending on resource availability, the patient then transitions between the three states *G&A*, *CC* or G^* until she eventually reaches the corresponding *Last Week* state, after which she is discharged from hospital (*Recovered* or *Dead*). We assume that a patient in a *Last Week* state only consumes half of the hospital resources, which mimics a half-week stay at the hospital. The inclusion of designated *Last Week* states allows us to account for the empirical fact that for some disease types, a large fraction of the patients require hospitalization for a few days only.

5 Numerical Experiments

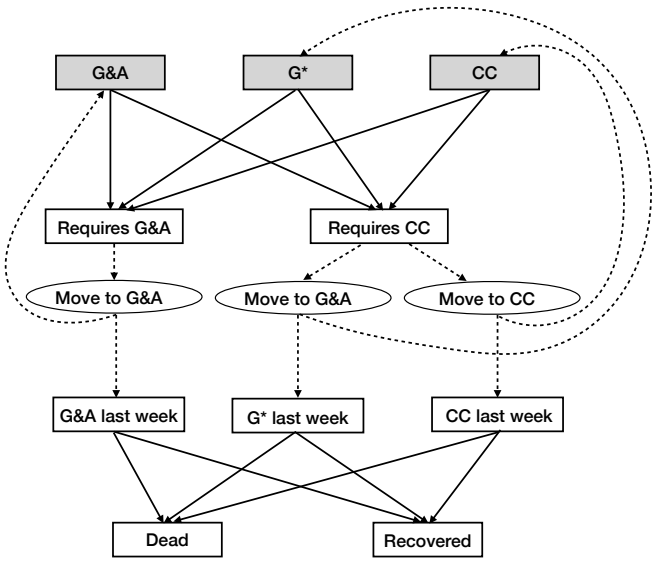
We next present the numerical results of our NHS England case study (Section 5.1) and compare our optimized schedule (hereafter OS) against a COVID prioritization policy (hereafter CP) that resembles the one implemented in England during the pandemic (Section 5.2). All experiments were run on a 2.7GHz quad-core Intel i7 processor with 16GB RAM using IBM ILOG CPLEX Optimization Studio 20.1. The runtimes of the fluid LP (3) ranged between 1.5 and 2 minutes.



(a)



(b)



(c)

Figure 8. Schematic representation of a patient DP, from admission to hospital (a) to discharge (c). Rectangular (oval) nodes correspond to states (actions), and grey shaded nodes are exploded in the subsequent subfigure. Dotted lines correspond to immediate actions and instantaneous transitions, and full lines represent weekly transitions.

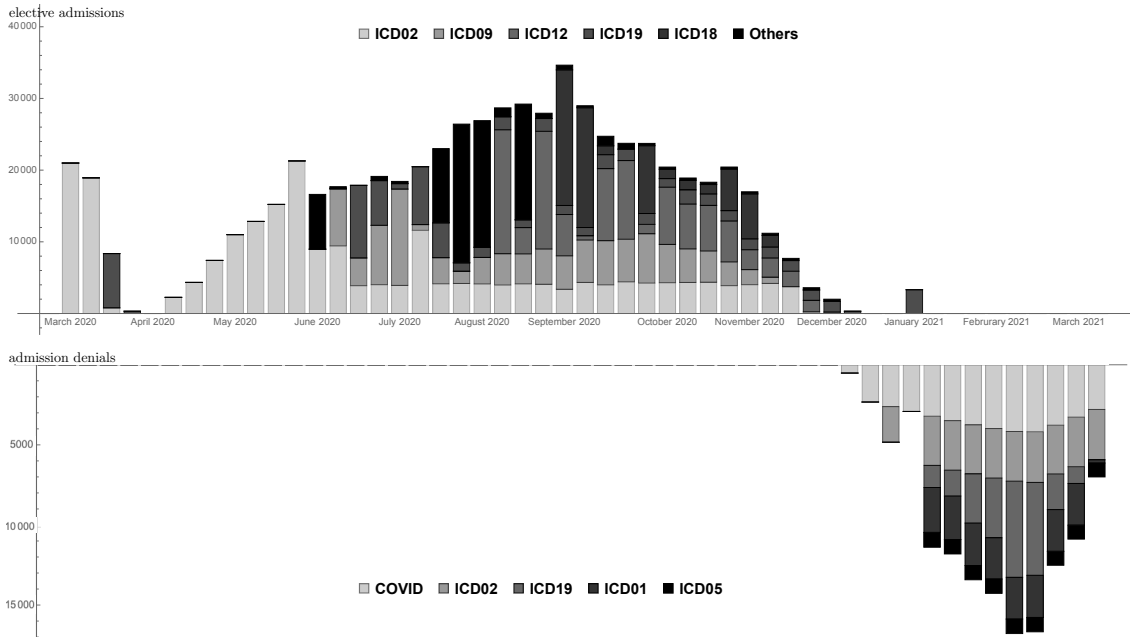


Figure 9. Weekly elective admissions (top) and admission denials (bottom) under the OS, categorized by disease group (ICD01: certain infectious and parasitic diseases; ICD02: neoplasms; ICD05: mental and behavioural disorders; ICD09: diseases of the circulatory system; ICD12: diseases of the skin and subcutaneous tissue; ICD18: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; ICD19: injury, poisoning and certain other consequences of external causes).

5.1 Optimized Schedule

Figure 9 shows the weekly elective admissions and emergency denials of the OS. Over the 52-week planning horizon, the OS admits to hospital 6,939,573 emergency patients (not shown), while 654,308 elective patients are admitted to hospital from week 1 to week 43 (upper part of Figure 9). Among the admitted elective patients, the most numerous groups are cancer patients (230,928), patients affected by diseases of the circulatory system (107,048) and diseases of the skin and subcutaneous tissue (101,790). In the last weeks of the planning horizon, due to the high inflow of COVID patients during the second wave of the pandemic, the available resources are insufficient to cope with the surge in demand. As a result, admission to hospital is denied to 125,346 emergency patients during weeks 38-52 (lower part of Figure 9). All of these patients are above 65 years of age, and most of them are COVID (40,962), cancer (30,069) and injury & poisoning (24,870) patients. The OS denies admission to hospital to these elderly patients as they have the lowest chances of benefiting from care.

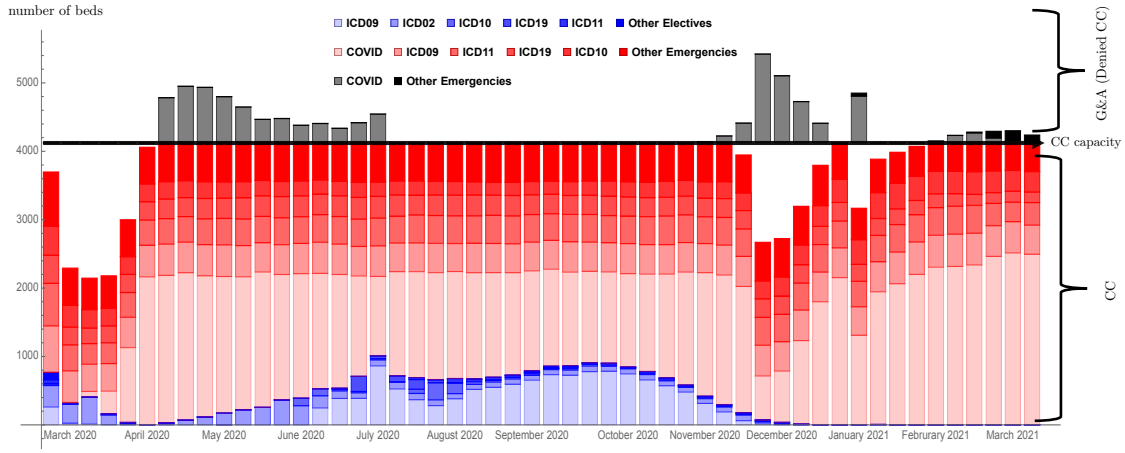


Figure 10. Weekly bed occupancy in CC by disease group (ICD02: neoplasms; ICD09: diseases of the circulatory system; ICD10: diseases of the respiratory system; ICD11: diseases of the digestive system; ICD19: injury, poisoning and certain other consequences of external causes). Patients who are denied CC, and have hence been moved to the $G^\star$ state, are shown above the CC capacity line.

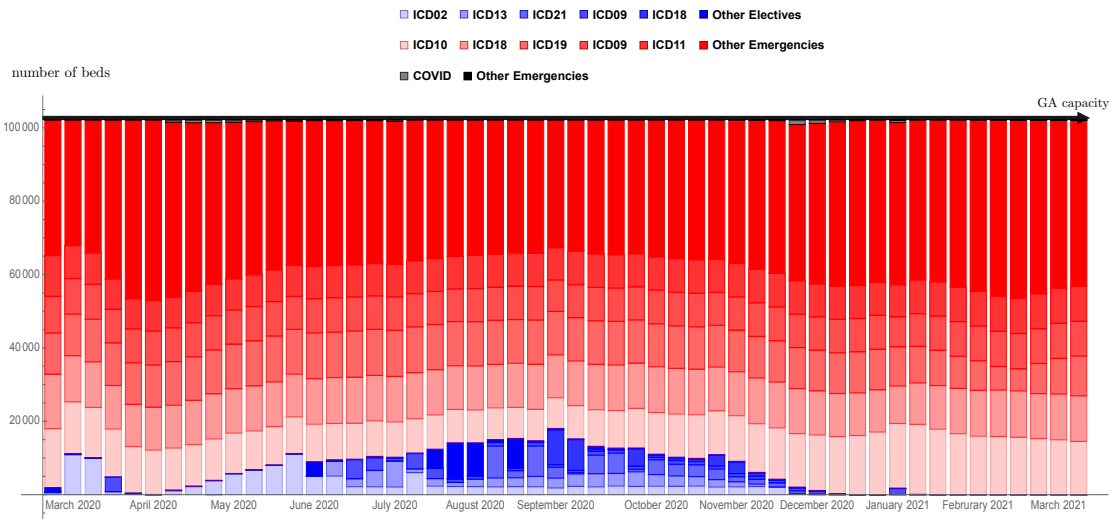


Figure 11. Weekly bed occupancy in G&A by disease group (ICD02: neoplasms; ICD09: diseases of the circulatory system; ICD10: diseases of the respiratory system; ICD11: diseases of the digestive system; ICD13: diseases of the musculoskeletal system and connective tissue; ICD18: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; ICD19: injury, poisoning and certain other consequences of external causes; ICD21: factors influencing health status and contact with health services).

Figures 10 and 11 show the bed occupancy in CC and G&A, respectively. A large share of CC beds is occupied by COVID patients (40.2% average occupancy over the planning horizon),

		Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
LP	G&A	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	CC	63%	93%	100%	100%	100%	100%	100%	100%	100%	76%	91%	99%	100%
DR	G&A	100%	101%	100%	100%	100%	100%	100%	100%	100%	100%	99%	101%	100%
	CC	63%	94%	99%	100%	100%	100%	100%	98%	99%	78%	93%	101%	101%
R	G&A	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
	CC	66%	95%	107%	108%	101%	104%	97%	96%	100%	76%	90%	100%	101%

Table 2. Average monthly G&A and CC bed occupancy for the fluid LP (LP), the deterministic rounded policy (DR) and the randomized policy (R). All numbers are rounded to the closest integer. The first 10 months (last 3 months) fall into the year 2020 (2021).

while 9.4% of the available CC beds are assigned (on average) to elective patients. During both the first and the second wave of the pandemic, capacity is insufficient, and CC is denied to some patients. The affected patients are almost exclusively COVID patients above 65 years of age, who are transferred to $G^\star$ due to their longer hospital stays as well as their lower capacity to benefit from treatment. G&A resources are fully utilized over the entire planning horizon, and the share of elective patients in G&A is on average 6.6%. The largest group of emergency patients in G&A are patients affected by diseases of the respiratory system, while cancer patients are the largest elective patient group. Overall, the OS results in 8,233,216 YLL over the 52 weeks planning horizon, with COVID patients contributing to 64% of the YLL.

We now investigate the performance of the deterministic rounded and the randomized policy from Section 3.3. Recall that both of these approximation schemes generate policies $\pi \in \Pi$ for the weakly coupled counting DP whose objective values are close to the objective value of the fluid LP (which itself overestimates the optimal value of the weakly coupled counting DP) and whose resource violations are small, with high probability. For our case study, the deterministic rounded policy results in a YLL increase of 0.02% (from 8,233,216 to 8,235,198) as well as a 0.04% higher G&A and a 0.31% higher CC occupancy (across the entire time horizon). Likewise, the randomized policy results in a YLL *de*crease of 0.01% (from 8,233,216 to 8,232,570) as well as a 0.05% higher G&A and a 1.56% higher CC occupancy. Table 2 compares the monthly bed occupancy of the fluid LP (3) with the bed occupancy of our two approximation schemes in further detail. We conclude that both policies offer approximations of high quality.

Figure 12 visualizes the trade-off between the competing objectives of minimizing YLL and costs. To this end, we define the rewards of the individual DPs as convex combinations of the patient's contributions to the overall YLL and cost of care. When the OS solely minimizes YLL, the total

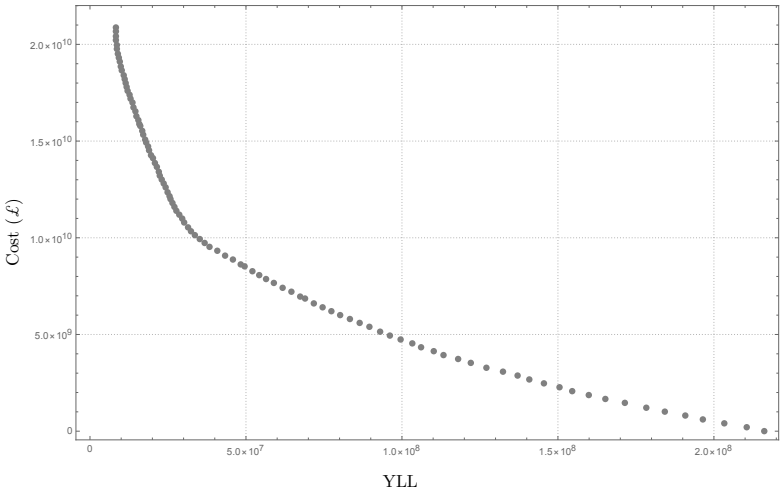


Figure 12. Pareto analysis: YLL vs. total cost of care.

healthcare costs amount to £20,659 million. If we were to (hypothetically) not treat any patients, on the other hand, the YLL would increase from 8,233,216 to 216,255,399. As expected, significant reductions of YLL (or total cost of care) can be achieved when optimizing convex combinations of both objectives, rather than one of them in isolation. Recall that our model focuses on the variable costs of treatments, which are directly proportional to the number of hospital admissions. Thus, treating no patients results in zero costs in Figure 12 since we disregard the fixed costs associated with the running and maintenance of the hospitals, staff, prevention activities and primary care.

5.2 Comparison with COVID Prioritization Policies

The results from the previous section suggest that denying hospital or CC admission to COVID patients might be beneficial in case of capacity shortages. This contrasts with current practice, where many countries prioritize COVID patients to the detriment of other patients. In the following, we thus compare our OS against a CP policy that always admits COVID patients and that strongly penalizes CC denial to COVID patients in the objective. Other than that, the CP policy coincides with the OS; in particular, within the aforementioned restrictions, the CP policy optimally schedules care across all patients groups.

Figure 13 shows that, while the total number of elective admissions is similar to the OS (655,415), emergency admission denials are significantly higher under the CP policy (153,092, +22.1% compared to the OS). Specifically, while all COVID patients are admitted to hospital,

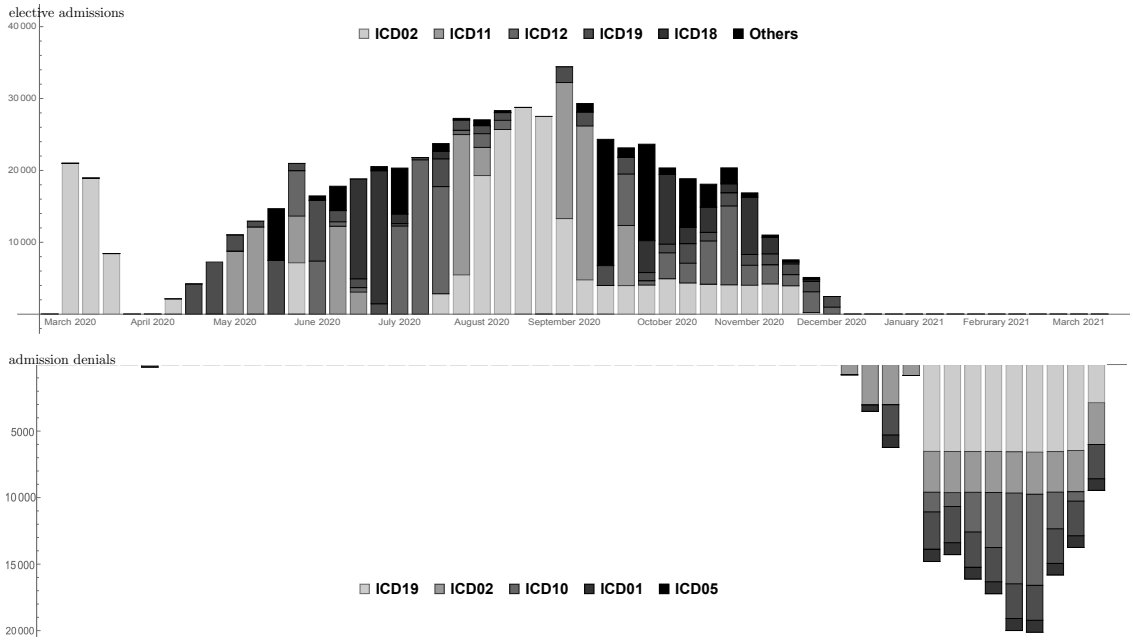


Figure 13. Weekly elective admissions (top) and admission denials (bottom) under the CP policy, categorized by disease group (ICD01: certain infectious and parasitic diseases; ICD02: neoplasms; ICD05: mental and behavioural disorders; ICD10: diseases of the respiratory system; ICD11: diseases of the digestive system; ICD12: diseases of the skin and subcutaneous tissue; ICD18: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; ICD19: injury, poisoning and certain other consequences of external causes).

emergency admission is denied to patients above 65 years of age affected by injury & poisoning (55,130), cancer (35,663) and diseases of the respiratory system (26,784). The higher numbers of emergency admission denials are due to the longer treatment of COVID patients, relative to patients affected by other diseases. Under the CP policy, an average 71.6% of the CC beds are occupied by COVID patients (+75.6% compared to the OS), and the CC occupancy reaches 100% during the second wave of the pandemic (weeks 42-52). The share of electives in CC and G&A is reduced to 4.4% (-53.2% compared to the OS) and 6.5% (-1.5% compared to the OS), respectively.

Overall, the prioritization of COVID patients in admission to hospital and CC leads to an 8.7% increase in the total YLL under the CP policy compared to the OS. Figure 14 shows a breakdown of this total 719,868 YLL across the different disease groups. Significant losses in years of life are seen for patients affected by injury & poisoning (318,955), diseases of the respiratory system (259,012), diseases of the circulatory system (108,085), diseases of the digestive system (85,134) and cancer (78,464), to the benefit of elderly COVID patients (275,691).

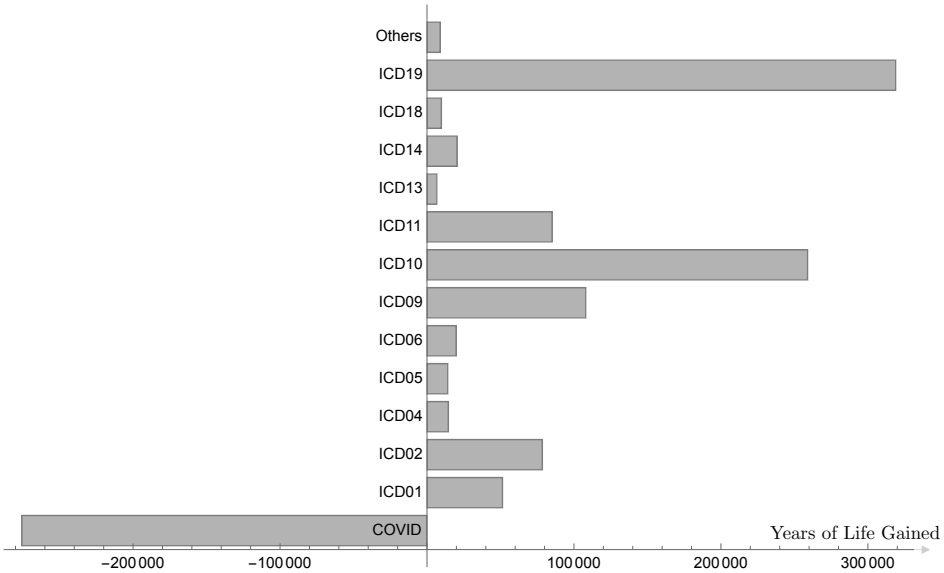


Figure 14. Years of Life Gained (*i.e.*, YLL avoided) by the OS relative to the CP policy, categorized by disease group (ICD01: certain infectious and parasitic diseases; ICD02: neoplasms; ICD04: endocrine, nutritional and metabolic diseases; ICD05: mental and behavioural disorders; ICD06: diseases of the nervous system; ICD09: diseases of the circulatory system; ICD10: diseases of the respiratory system; ICD11: diseases of the digestive system; ICD13: diseases of the musculoskeletal system and connective tissue; ICD14: diseases of the genitourinary system; ICD18: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; ICD19: injury, poisoning and certain other consequences of external causes).

We emphasize that the CP policy constitutes an overly optimistic representation of the current practice in England, where not only COVID patients are prioritized but also the other patient groups are scheduled suboptimally based on static prioritization schemes. Thus, we expect our results to underestimate the benefits of the OS over the current practice. We refer to the accompanying paper D'Aeth et al. (2021) for a comparison of the OS against a set of government admission policies across a range of scenarios.

6 Extensions

Our approach of modeling the health system via a weakly coupled counting DP and subsequently determining a near-optimal solution via the fluid LP (3) is very versatile. In this section, we highlight some extensions of our method that can help to obtain better informed, fairer and more

resilient decisions as well as further insights into the characteristics of the optimal solution.

Relaxation of Model Assumptions. Our healthcare model makes a number of strong assumptions that can be relaxed. Firstly, with the exception of the patients that have been denied CC (and that have thus transferred to the state $G^\star$), the transitions in our model—such as the weekly probability of an elective patient turning into an emergency or the weekly probability of a hospitalized patient recovering or dying—are Markovian and hence memoryless. In reality, disease progression may exhibit a more complicated dependence on waiting time, and this time dependency differs across the different patient groups. We can readily model non-Markovian transitions by adding memory to the states of the patient DPs (*e.g.*, how many weeks has a patient been waiting for her surgery, and how many weeks has a patient spent in hospital). An interesting problem that arises in this context is how to best approximate non-Markovian transitions via a small number of additional states. Secondly, our model assumes that the timing and the magnitude of the patient inflows as well as the availability of staff is independent of the hospital occupancy rates. Since COVID is highly infectious, however, both non-COVID patients and hospital staff get exposed to the virus and—in absence of protective behavior—may spread it to the community. It would therefore be instructive to study the impact of hospital occupancy rates, which are immediate consequences of the admissions decisions in our model, on hospital acquired infections, changes in care seeking behavior as well as workload-dependent staff absenteeism (Green et al., 2013). Thirdly, our model assumes that capacity can be re-assigned between COVID and non-COVID cases on short notice. Since COVID patients require isolation and dedicated staff to reduce the risk of infections, this is not the case in practice, and our model may therefore overestimate the available capacity. We believe that this issue is attenuated by the fact that we are modeling the health system of an entire nation, as opposed to an individual hospital. Finally, our model disregards geographical differences in patient numbers, hospital resources and treatment efficiency. While this appears to be an acceptable approximation in our case study, a more elaborate model could subdivide the country into different regions and impose that patients can only be treated in hospitals that are sufficiently close.

Alternative Objectives, Constraints and Decisions. While YLL and costs are natural objectives to minimize, one could also consider the incorporation of inequity aversion in the population distribution of healthcare utilization and/or health outcomes. . This would enable the policy

maker to sacrifice some efficiency in favor of providing people of different age, gender, ethnicity and medical history equal chances of survival. Further refinements of our model could include additional policy restrictions, such as prioritizing CC access for patients that are already hospitalized or ensuring that patients of every disease and age group are admitted to hospital within a certain maximum time interval. While some of these restrictions can be readily included as linear constraints, others may require the imposition of logical constraints and thus result in mixed-integer linear programs. Finally, additional tactical and strategic decisions, such as the construction of temporary field hospitals, the enlistment of retired staff and medical students, changes to the employed staff-to-bed ratios as well as alterations in the design of the hospitals (*e.g.*, the erection of isolation wards for COVID patients), can be readily incorporated into our model through the inclusion of additional continuous or discrete decision variables.

Regularization against Data Uncertainty. Our healthcare model can be safeguarded against the impact of uncertainty in the patient inflows, transition probabilities and resource availabilities. To this end, we can replace the weakly coupled counting DP with a robust counterpart that determines the optimal policy in view of the worst rewards and transition probabilities from within a pre-specified uncertainty set, which can itself be selected so as to offer rigorous statistical guarantees. Robust policies have attracted significant interest in the context of Markov decision processes (Iyengar, 2005; Nilim and Ghaoui, 2005; Wiesemann et al., 2013), and similar concepts can be readily applied to our weakly coupled counting DP. Assuming that the uncertainty set is polyhedral, the resulting robust version of our fluid LP (3) is amenable to the 'robust optimization trick' (Ben-Tal et al., 2009; Bertsimas et al., 2011) and thus reduces to a linear program of moderately larger size than the nominal fluid LP (3).

Sensitivity Analysis. An important advantage of our LP-based approach is that the optimal solution to our healthcare model is amenable to sensitivity analysis. The shadow prices of the patient inflow constraints, for example, allow us to evaluate the impact of additional elective and emergency patients of a particular age group and disease type at different times during the pandemic. The shadow prices of the resource constraints inform about the value of different resources over time, and they allow to investigate the impact of changes to the required staff-to-bed ratios. The sensitivity of the optimal objective value with respect to the transition probabilities, finally, allows us to quantify the impact of improvements to certain treatments (*e.g.*, the administration

of medicines to shorten the hospitalization of COVID patients) on the overall health outcome.

References

Adelman, D. and A. J. Mersereau (2008). Relaxations of weakly coupled stochastic dynamic programs. *Operations Research 56*(3), 712–727.

Argenziano, M., K. Fischkoff, and C. R. Smith (2020). Surgery scheduling in a crisis. *New England Journal of Medicine 382*(23), e87.

Bekker, R. and P. M. Koeleman (2011). Scheduling admissions and reducing variability in bed demand. *Health Care Management Science 14*(3), 237–249.

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America 38*(8), 716–719.

Ben-Tal, A., L. El Ghaoui, and A. Nemirovski (2009). *Robust Optimization*. Princeton University Press.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control*, Volume 1. Athena Scientific.

Bertsekas, D. P. and J. Tsitsiklis (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Bertsimas, D., D. B. Brown, and C. Caramanis (2011). Theory and applications of robust optimization. *SIAM Review 53*(3), 464–501.

Bertsimas, D., G. Lukin, L. Mingardi, et al. (2020). COVID-19 mortality risk assessment: An international multi-center study. *PLOS ONE 15*(12), 1–13.

Bertsimas, D. and V. V. Mišić (2016). Decomposable Markov decision processes: A fluid optimization approach. *Operations Research 64*(6), 1537–1555.

Bertsimas, D., J. Pauphilet, J. Stevens, et al. (2020). Predicting inpatient flow at a major hospital using interpretable analytics. *medRxiv 2020.05.12.20098848v2*.

Boutilier, C., R. Dearden, and M. Goldszmidt (1995). Exploiting structure in policy construction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Volume 14, pp. 1104–1113.

Burki, T. K. (2020). Cancer guidelines during the COVID-19 pandemic. *The Lancet Oncology 21*(5), 629–630.

Chan, C. W., V. F. Farias, N. Bambos, et al. (2012). Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research 60*(6), 1323–1341.

Christen, P., J. D'Aeth, A. Lochen, et al. (2021). The J-IDEA pandemic planner: A framework for implementing hospital provision interventions during the COVID-19 pandemic. *Medical Care, Published Ahead-of-Print*.

D'Aeth, J., S. Ghosal, F. Grimm, et al. (2021). Who shall live? Optimal scheduling rules for elective care to minimize years of life lost during the SARS-CoV-2 pandemic: an application to England. *Under Review*.

Davis, C., M. Gao, M. Nichols, et al. (2020). Predicting hospital utilization and inpatient mortality of patients tested for COVID-19. *medRxiv 2020.12.04.20244137*.

De Farias, D. P. and B. Van Roy (2004). On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research 29*(3), 462–478.

Déry, J., A. Ruiz, F. Routhier, et al. (2020). A systematic review of patient prioritization tools in non-emergency healthcare services. *Systematic Reviews 9*(227), 1–14.

DIVI (2020). Entscheidungen über die Zuteilung von Ressourcen in der Notfall und der Intensivmedizin im Kontext der COVID-19-Pandemie. `https://www.divi.de/joomlatools-files/docman-files/publikationen/covid-19-dokumente/200325-covid-19-ethik-empfehlung-v1.pdf`. Accessed on 6 February 2021.

Eichberg, D. G., A. H. Shah, E. M. Luther, et al. (2020). Letter: Academic neurosurgery department response to COVID-19 pandemic: the University of Miami/Jackson Memorial Hospital model. *Neurosurgery 87*(1), E63–E65.

Fujita, K., T. Ito, Z. Saito, et al. (2020). Impact of COVID-19 pandemic on lung cancer treatment scheduling. *Thoracic Cancer 11*(10), 2983–2986.

Gao, Y., G.-Y. Cai, W. Fang, et al. (2020). Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nature Communications 11*(1), 1–10.

Gardner, T., C. Fraser, and S. Peytrignet (2020). Elective care in England: Assessing the impact of COVID-19 and where next. Technical report, The Health Foundation.

Gittins, J., K. Glazebrook, and R. Weber (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.

Green, L. V., S. Savin, and N. Savva (2013). "Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science 59*(10), 2237–2256.

Guestrin, C., D. Koller, and R. Parr (2001). Multiagent planning with factored MDPs. In *Advances in Neural Information Processing Systems*, Volume 1, pp. 1523–1530.

Guestrin, C., D. Koller, R. Parr, et al. (2003). Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research 19*, 399–468.

Haw, D., P. Christen, G. Forchini, et al. (2020). DAEDALUS: An economic-epidemiological model to optimize economic activity while containing the SARS-CoV-2 pandemic. Technical report, Imperial College London.

Hawkins, J. T. (2003). *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. Ph. D. thesis, Massachusetts Institute of Technology.

Helm, J. E., S. AhmadBeygi, and M. P. Van Oyen (2011). Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management 20*(3), 359–374.

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research 30*(2), 257–280.

Joebges, S. and N. Biller-Andorno (2020). Ethics guidelines on COVID-19 triage—an emerging international consensus. *Critical Care 24*(1), 201.

Kim, S.-H., C. W. Chan, M. Olivares, et al. (2015). ICU Admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science 61*(1), 19–38.

MacCormick, A. D., W. G. Collecutt, and B. R. Parry (2003). Prioritizing patients for elective surgery: A systematic review. *ANZ Journal of Surgery 73*(8), 633–642.

McCabe, R., N. Schmit, P. Christen, et al. (2020). Adapting hospital capacity to meet changing demands during the COVID-19 pandemic. *BMC Medicine 18*(329), 1–12.

Meng, F., J. Qi, M. Zhang, et al. (2015). A robust optimization model for managing elective admission in a public hospital. *Operations Research 63*(6), 1452–1467.

Moris, D. and E. Felekouras (2020). Surgery scheduling in a crisis: Effect on cancer patients. *Journal of BUON 25*(4), 2123–2124.

Negopdiev, D., C. Collaborative, and E. Hoste (2020). Elective surgery cancellations due to the COVID-19 pandemic: global predictive modelling to inform surgical recovery plans. *British Journal of Surgery 107*(11), 1440–1449.

NHS (2020a). Bed Availability and Occupancy. https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/. Accessed on 6 February 2021.

NHS (2020b). COVID-19 Hospital Activity. https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-hospital-activity/. Accessed on 6 February 2021.

NHS (2020c). Critical Care Bed Capacity and Urgent Operations Cancelled. https://www.england.nhs.uk/statistics/statistical-work-areas/critical-care-capacity/. Accessed on 6 February 2021.

NHS (2020d). Important and Urgent: Next Steps on NHS Response to COVID-19. https://www.england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2020/03/urgent-next-steps-on-nhs-response-to-covid-19-letter-simon-stevens.pdf. Accessed on 6 February 2021.

NHS Digital (2020). Hospital Episode Statistics (HES). https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics. Accessed on 6 February 2021.

NHS England (2020). NHS England National Cost Collection for the NHS. https://www.england.nhs.uk/national-cost-collection/. Accessed on 6 February 2021.

NICE (2020). COVID-19 Rapid Guideline: Critical care in adults. https://www.nice.org.uk/guidance/ng159. Accessed on 6 February 2021.

Nilim, A. and L. E. Ghaoui (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research 53*(5), 780–798.

Office for National Statistics (2019). Past and projected period and cohort life tables, 2018-based, UK. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/pastandprojecteddatafromtheperiodandcohortlifetables/1981to2068. Accessed on 6 February 2021.

Ouyang, H., N. T. Argon, and S. Ziya (2020). Allocation of intensive care unit beds in periods of high demand. *Operations Research 68*(2), 591–608.

Perez-Guzman, P. N., A. Daunt, S. Mukherjee, et al. (2020). Clinical characteristics and predictors of outcomes of hospitalized patients with COVID-19 in a multi-ethnic London NHS Trust: A retrospective cohort study. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America ciaa1091*.

Phua, J., L. Weng, L. Ling, et al. (2020). Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. *The Lancet Respiratory Medicine 8*(5), 506–517.

Powell, W. B. (2007). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

RCP London (2020). Safe medical staffing. https://www.rcplondon.ac.uk/projects/outputs/safe-medical-staffing. Accessed on 6 February 2021.

Riccioni, L., G. Bertolini, A. Giannini, et al. (2020). Raccomandazioni di etica clinica per l'ammissione a trattamenti intensivi e per la loro sospensione, in condizioni eccezionali di squilibrio tra necessità e risorse disponibili. *Recenti Progressi in Medicina 111*(4), 207–211.

Rizmie, D., M. Miraldo, R. Atun, et al. (2019). The effect of extreme temperature on emergency admissions across vulnerable populations in England: An observational study. *The Lancet 394*(Special Issue 2), S7.

Rockafellar, R. T. and R. J.-B. Wets (1997). *Variational Analysis*. Springer.

Royal College of Nursing (2020). Staffing levels — Advice guides — Royal College of Nursing. `https://www.rcn.org.uk/get-help/rcn-advice/staffing-levels`. Accessed on 6 February 2021.

Shi, P., J. Helm, J. Deglise-Hawkinson, et al. (2019). Timing it right: Balancing inpatient congestion versus readmission risk at discharge. *Available at SSRN 3202975*.

Soltany, A., M. Hamouda, A. Ghzawi, et al. (2020). A scoping review of the impact of COVID-19 pandemic on surgical practice. *Annals of Medicine and Surgery 57*, 24–36.

Sud, A., M. E. Jones, J. Broggio, et al. (2020). Collateral damage: the impact on outcomes from cancer surgery of the COVID-19 pandemic. *Annals of Oncology 31*(8), 1065–1074.

The Nuffield Council on Bioethics (2020). Statement: The need for national guidance on resource allocation decisions in the COVID-19 pandemic. `https://www.nuffieldbioethics.org/news/statement-the-need-for-national-guidance-on-resource-allocation-decisions-in-the-covid-19-pandemic`. Accessed on 6 February 2021.

Tzeng, C.-W. D., M. Teshome, M. H. Katz, et al. (2020). Cancer surgery scheduling during and after the COVID-19 first wave: the MD Anderson cancer center experience. *Annals of Surgery 272*(2), e106–e111.

Vaid, A., S. Somani, A. Russak, et al. (2020). Machine learning to predict mortality and critical events in COVID-19 positive New York City patients: A cohort study. *Journal of Medical Internet Research 22*(11), 1–19.

Wiesemann, W., D. Kuhn, and B. Rustem (2013). Robust Markov decision processes. *Mathematics of Operations Research 38*(1), 153–183.

Yoon, D. H., S. Koller, P. M. N. Duldulao, et al. (2020). COVID-19 impact on colorectal daily practice—how long will it take to catch up? *Journal of Gastrointestinal Surgery 25*, 260–268.

Appendix: Proofs

Proof of Proposition 1. Denote by $\tilde{\sigma} = \{\tilde{\sigma}_t\}_t$ the random state evolution of the counting DP under policy π and by $\tilde{s}' = \{\tilde{s}'_t\}_t$ the random state evolution of the weakly coupled DP under any fixed policy π' satisfying the conditions in the statement of the proposition, respectively. We also define the counting state evolution of the weakly coupled DP as $\tilde{\sigma}' = \{\tilde{\sigma}'_t\}_t$ with

$$\tilde{\sigma}'_t(s) \;=\; \big|\{i \in \mathcal{I} \,:\, \tilde{s}'_{ti} = s\}\big| \qquad \forall t \in \mathcal{T},\; \forall s \in \mathcal{S}.$$

We prove the statement in two steps. We first argue that the counting state evolutions $\tilde{\sigma}_t$ and $\tilde{\sigma}'_t$ of the counting DP and the weakly coupled DP share the same distributions for all $t \in \mathcal{T}$. We subsequently use this insight to show that the expected total rewards of π and π' in their respective DPs coincide. Since the weakly coupled DP has no resource constraints, the policy π' is feasible by construction, and the statement of the proposition thus follows.

As for the first step, we note that the conditions in the statement of the proposition ensure that whenever the counting states $\tilde{\sigma}_t$ and $\tilde{\sigma}'_t$ of the counting DP and the weakly coupled DP are both equal to $\sigma_t \in \mathfrak{S}$, then the policies π_t and π'_t apply each action $a \in \mathcal{A}$ to DPs in state $s \in \mathcal{S}$ precisely $[\pi_t(\sigma_t)](s,a)$ many times. Moreover, Definition 3 implies that the transition probabilities $\mathfrak{p}$ of the counting DP record the aggregate transitions of n i.i.d. DPs with individual transition probabilities p under policy π, whereas the weakly coupled DP by construction records the individual transitions of these DPs under policy π'. We thus conclude that the transition probabilities of the counting states $\tilde{\sigma}_t$ and $\tilde{\sigma}'_t$ are identical under π and π'. Moreover, the construction of the initial state distribution $\mathfrak{q}$ in Definition 3 and the definition of $\tilde{s}'_{1i}$, $i \in \mathcal{I}$, as i.i.d. random variables governed by the distribution q imply that $\tilde{\sigma}_1$ and $\tilde{\sigma}'_1$ share the same distribution. A simple induction therefore shows that $\tilde{\sigma}_t$ and $\tilde{\sigma}'_t$ share the same distributions across all time periods $t \in \mathcal{T}$.

In view of the second step, we observe that the expected total reward of the weakly coupled

DP under policy π' evaluates to

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{i=1}^{n} r_t(\tilde{s}'_{ti},\pi'_{ti}(\tilde{s}'_{ti}))\right] &= \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{i=1}^{n}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r_t(s,a)\cdot\mathbf{1}\left[\tilde{s}'_{ti}=s\ \wedge\ \pi'_{ti}(\tilde{s}'_{ti})=a\right]\right] \\
&= \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r_t(s,a)\cdot\left|\{i\in\mathcal{I}\,:\,\tilde{s}'_{ti}=s\ \wedge\ \pi'_{ti}(\tilde{s}'_{ti})=a\}\right|\right] \\
&= \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r_t(s,a)\cdot[\pi_t(\tilde{\sigma}'_t)](s,a)\right] \\
&= \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r_t(s,a)\cdot[\pi_t(\tilde{\sigma}_t)](s,a)\right],
\end{aligned}$$

where the first and second identity hold by construction, the third identity uses the properties of π' from the statement of the proposition, and the last identity exploits the fact that $\tilde{\sigma}_t$ and $\tilde{\sigma}'_t$ share the same distribution. The statement now follows from the fact that the last expression evaluates the expected total reward of the counting DP under the policy π. □

Proof of Proposition 2. In view of statement *(i)*, we first verify that $\overline{\pi}\in\overline{\Pi}^{\mathrm{C}}$, that is, $\overline{\pi}$ is a feasible policy for the fluid DP. To this end, we need to confirm that $\overline{\pi}_t(\sigma)\in\overline{\mathfrak{A}}^{\mathrm{C}}_t(\sigma)$ for all $t\in\mathcal{T}$ and all $\sigma\in\overline{\mathfrak{S}}$. Due to the feasibility of $\overline{\pi}^0$, this is the case for all $t\in\mathcal{T}$ and all $\sigma\in\overline{\mathfrak{S}}\backslash\{\sigma^{\mathrm{LP}}_t\}$. To verify that $\overline{\pi}_t(\sigma^{\mathrm{LP}}_t)\in\overline{\mathfrak{A}}^{\mathrm{C}}_t(\sigma^{\mathrm{LP}}_t)$ for all $t\in\mathcal{T}$ as well, we observe that $[\overline{\pi}_t(\sigma^{\mathrm{LP}}_t)]_j\in\overline{\mathfrak{A}}_j$, $j\in\mathcal{J}$, by construction, while $\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot[\overline{\pi}_t(\sigma^{\mathrm{LP}}_t)])_j(s,a)\leqslant b_{tl}$, $\sum_{a\in\mathcal{A}_j}[\overline{\pi}_t(\sigma^{\mathrm{LP}}_t)]_j(s,a)=\sigma^{\mathrm{LP}}_{tj}(s)$ and $[\overline{\pi}_t(\sigma^{\mathrm{LP}}_t)]_j(s,a)=0$ hold for all $j\in\mathcal{J}$, $s\in\mathcal{S}_j$, $a\in\mathcal{A}_j\backslash\mathcal{A}_{jt}(s)$, $l\in\mathcal{L}$ and $t\in\mathcal{T}$ due to the third, fourth and fifth constraint in the fluid LP (3), respectively.

We next show via induction over $t\in\mathcal{T}$ that the random state evolution $\tilde{\sigma}=\{\tilde{\sigma}_t\}_t$ of the fluid DP under the policy $\overline{\pi}$ satisfies $\tilde{\sigma}=\sigma^{\mathrm{LP}}$ almost surely. For $\tilde{\sigma}_1$, this immediately follows from the initial state probabilities in the definition of the fluid DP as well as the first constraint of the fluid LP. Assume now that $\tilde{\sigma}_t=\sigma^{\mathrm{LP}}_t$ almost surely for some $t\in\mathcal{T}\backslash\{T\}$. We then have $\overline{\pi}_t(\tilde{\sigma}_t)=\pi^{\mathrm{LP}}_t$ almost surely, and the transition probabilities in the definition of the fluid DP as well as the second constraint of the fluid LP imply that $\tilde{\sigma}_{t+1}=\sigma^{\mathrm{LP}}_{t+1}$ almost surely as well. Since $\tilde{\sigma}=\sigma^{\mathrm{LP}}$ almost surely and $\overline{\pi}_t(\sigma^{\mathrm{LP}}_t)=\pi^{\mathrm{LP}}_t$ for all $t\in\mathcal{T}$, the expected total reward of the policy $\overline{\pi}$ is indeed θ^{LP} as claimed. This proves the first statement.

Consider now statement *(ii)*. A similar induction argument as in the previous paragraph shows

that the random state evolution $\tilde{\sigma} = \{\tilde{\sigma}_t\}_t$ of the fluid DP satisfies $\tilde{\sigma} = \sigma^{\mathrm{LP}}$ and the policy $\overline{\pi}$ of the fluid DP satisfies $\overline{\pi}_t(\tilde{\sigma}_t) = \pi_t^{\mathrm{LP}}$ for all $t \in \mathcal{T}$ almost surely. The feasibility of $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ in the fluid LP (3) then follows from the initial state and transition probabilities in the definition of the fluid DP as well as the fact that $\overline{\pi}$ is a feasible policy for the fluid DP. Moreover, since $\pi_t^{\mathrm{LP}} = \overline{\pi}_t(\tilde{\sigma}_t)$ almost surely for all $t \in \mathcal{T}$, the objective value of $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ in (3) is indeed θ^{DP} as claimed. $\square$

Proof of Theorem 1. Denote by $F(\tau, \sigma_\tau)$ the truncated fluid LP that starts in period $\tau \in \mathcal{T}$ with the state $\sigma_\tau \in \overline{\mathfrak{S}}$:

$$
\begin{array}{lll}
\underset{\sigma,\, \pi}{\text{maximize}} & \displaystyle\sum_{t=\tau}^{T} f(t, \pi_t) & \\
\text{subject to} & \sigma_{t+1,j}(s') = \displaystyle\sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} p_{jt}(s' \,|\, s, a) \cdot \pi_{tj}(s, a) & \forall j \in \mathcal{J},\ \forall s' \in \mathcal{S}_j,\ \forall t = \tau, \ldots, T-1 \\
& \pi_t \in \overline{\Pi}_t^{\mathrm{1C}}(\sigma_t) & \forall t = \tau, \ldots, T \\
& \sigma_{tj}(s) \geqslant 0 & \forall j \in \mathcal{J},\ \forall s \in \mathcal{S}_j,\ \forall t = \tau, \ldots, T
\end{array}
$$

Here, the objective function satisfies

$$
f(t, \pi_t) \;=\; \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} r_{jt}(s, a) \cdot \pi_{tj}(s, a),
$$

and the constraints satisfy

$$
\pi_t \in \overline{\Pi}_t^{\mathrm{1C}}(\sigma_t) \quad \Longleftrightarrow \quad \left\{
\begin{array}{ll}
\displaystyle\sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} c_{tlj}(s, a) \cdot \pi_{tj}(s, a) \leqslant b_{tl} & \forall l \in \mathcal{L} \\
\displaystyle\sum_{a \in \mathcal{A}_j} \pi_{tj}(s, a) = \sigma_{tj}(s) & \forall j \in \mathcal{J},\ \forall s \in \mathcal{S}_j \\
\pi_{tj}(s, a) = 0 & \forall j \in \mathcal{J},\ \forall s \in \mathcal{S}_j,\ \forall a \in \mathcal{A}_j \backslash \mathcal{A}_{jt}(s) \\
\pi_{tj}(s, a) \geqslant 0 & \forall j \in \mathcal{J},\ \forall (s, a) \in \mathcal{S}_j \times \mathcal{A}_j.
\end{array}
\right.
$$

Setting $F(T+1, \sigma_{T+1}) = 0$ for all $\sigma_{T+1} \in \overline{\mathfrak{S}}$, one readily verifies that the truncated fluid LP satisfies

$$
F(t, \sigma_t) \;=\; \max_{\pi_t \in \overline{\Pi}_t^{\mathrm{1C}}(\sigma_t)} \left\{ f(t, \pi_t) + F\left(t+1,\ \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} p_{jt}(s' \,|\, s, a) \cdot \pi_{tj}(s, a) \right) \right\}.
$$

We show via induction that $F(t, \cdot)$ is concave for all $t \in \mathcal{T} \cup \{T+1\}$. This is trivially the case for

$t = T + 1$. Assume now that $F(t+1, \cdot)$ is concave and represent $F(t, \cdot)$ as

$$F(t, \sigma_t) \;=\; \max_{\pi_t \in \overline{\mathfrak{A}}} \left\{ f(t, \pi_t) + F\left(t+1,\; \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} p_{jt}(s' \,|\, s, a) \cdot \pi_{tj}(s, a) \right) - \mathbb{I}(\sigma_t, \pi_t) \right\},$$

where $\mathbb{I}(\sigma_t, \pi_t) = 0$ if $\pi_t \in \overline{\Pi}_t^{1\mathrm{C}}(\sigma_t)$ and $\mathbb{I}(\sigma_t, \pi_t) = \infty$ otherwise. Since $\mathbb{I}(\sigma_t, \pi_t)$ is convex, $F(t, \cdot)$ is concave as it represents a sup-projection of a concave function (Rockafellar and Wets, 1997, Proposition 2.22). We thus conclude that $F(t, \cdot)$ is concave for all $t \in \mathcal{T} \cup \{T+1\}$.

Denote now by $G(t, \sigma_t)$ the optimal value of the weakly coupled counting DP, which satisfies

$$G(t, \sigma_t) \;=\; \max_{\pi_t \in \Pi_t^{1\mathrm{C}}(\sigma_t)} \left\{ f(t, \pi_t) + \mathbb{E}\left[G(t+1, \tilde{\sigma}_{t+1}) \,|\, \tilde{\sigma}_{t+1} \sim \mathfrak{p}(\cdot \,|\, \sigma_t, \pi_t) \right] \right\}$$

for $t \in \mathcal{T}$ and $\sigma_t \in \mathfrak{S}$ with $\Pi_t^{1\mathrm{C}}(\sigma_t) = \overline{\Pi}_t^{1\mathrm{C}}(\sigma_t) \cap \mathfrak{A}$ as well as $G(T+1, \sigma_t) = 0$ for all $\sigma_t \in \mathfrak{S}$. We show via induction that $F(t, \sigma_t) \geqslant G(t, \sigma_t)$ for all $t \in \mathcal{T} \cup \{T+1\}$ and all $\sigma_t \in \mathfrak{S}$. The statement trivially holds for $t = T + 1$. Assume now that $F(t+1, \sigma_{t+1}) \geqslant G(t+1, \sigma_{t+1})$ for some $t \in \mathcal{T}$ and all $\sigma_{t+1} \in \mathfrak{S}$. We then have

$$\begin{aligned}
F(t, \sigma_t) \;&=\; \max_{\pi_t \in \overline{\Pi}_t^{1\mathrm{C}}(\sigma_t)} \left\{ f(t, \pi_t) + F\left(t+1,\; \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} p_{jt}(s' \,|\, s, a) \cdot \pi_{tj}(s, a) \right) \right\} \\
&\geqslant\; \max_{\pi_t \in \Pi_t^{1\mathrm{C}}(\sigma_t)} \left\{ f(t, \pi_t) + F\left(t+1,\; \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} p_{jt}(s' \,|\, s, a) \cdot \pi_{tj}(s, a) \right) \right\} \\
&=\; \max_{\pi_t \in \Pi_t^{1\mathrm{C}}(\sigma_t)} \left\{ f(t, \pi_t) + F\left(t+1,\; \mathbb{E}\left[\tilde{\sigma}_{t+1} \,|\, \tilde{\sigma}_{t+1} \sim \mathfrak{p}(\cdot \,|\, \sigma_t, \pi_t) \right] \right) \right\} \\
&\geqslant\; \max_{\pi_t \in \Pi_t^{1\mathrm{C}}(\sigma_t)} \left\{ f(t, \pi_t) + \mathbb{E}\left[F\left(t+1,\; \tilde{\sigma}_{t+1} \right) \,|\, \tilde{\sigma}_{t+1} \sim \mathfrak{p}(\cdot \,|\, \sigma_t, \pi_t) \right] \right\} \\
&\geqslant\; \max_{\pi_t \in \Pi_t^{1\mathrm{C}}(\sigma_t)} \left\{ f(t, \pi_t) + \mathbb{E}\left[G\left(t+1,\; \tilde{\sigma}_{t+1} \right) \,|\, \tilde{\sigma}_{t+1} \sim \mathfrak{p}(\cdot \,|\, \sigma_t, \pi_t) \right] \right\} \\
&=\; G(t, \sigma_t)
\end{aligned}$$

for all $\sigma_t \in \mathfrak{S}$, where the first and last identities hold by definition of F and G, respectively. The first inequality follows from $\Pi_t^{1\mathrm{C}}(\sigma_t) = \overline{\Pi}_t^{1\mathrm{C}}(\sigma_t) \cap \mathfrak{A} \subseteq \overline{\Pi}_t^{1\mathrm{C}}(\sigma_t)$, the second equality holds since the expected number of DPs in the j-th counting DP that are in state s' at time $t+1$ is the sum of all DPs in that counting DP that are in any state s at time t and whose associated action a transitions

them into state s', the individual probability of which is given by $p_{jt}(s' \,|\, s, a)$. The second inequality is due to Jensen's inequality, which is applicable since F has been shown to be concave, and the last inequality follows from the induction hypothesis.

Denote now by θ^{LP} the optimal objective value of the fluid LP (3) and by θ^{DP} the expected total reward of the weakly coupled counting DP under an optimal policy, respectively. We have

$$\begin{aligned}\theta^{\mathrm{LP}} \;=\; F(1,\,\sigma_1) \;&=\; F\left(1,\,\mathbb{E}\left[\tilde{\sigma}_1 \,|\, \tilde{\sigma}_1 \sim \mathfrak{q}\right]\right) \\ &\geqslant\; \mathbb{E}\left[F\left(1,\,\tilde{\sigma}_1\right) | \,\tilde{\sigma}_1 \sim \mathfrak{q}\right] \;\geqslant\; \mathbb{E}\left[G\left(1,\,\tilde{\sigma}_1\right) | \,\tilde{\sigma}_1 \sim \mathfrak{q}\right] \;=\; \theta^{\mathrm{DP}},\end{aligned}$$

where $\sigma_{1j} = n_j \cdot q_j$, $j \in \mathcal{J}$, and the two inequalities follow from Jensen's inequality and our induction argument from the previous paragraph, respectively. The claim of the theorem thus follows. □

Proof of Observation 1. To see that $\overline{\pi} \in \overline{\Pi}$, we need to show that $\overline{\pi}_t \in \overline{\Pi}_t$ for all $t \in \mathcal{T}$, that is, $\overline{\pi}_t(\sigma) \in \overline{\mathfrak{A}}_t(\sigma)$ for all $t \in \mathcal{T}$ and $\sigma \in \overline{\mathfrak{S}}$, which in turn amounts to showing that $[\overline{\pi}_t(\sigma)]_j \in \overline{\mathfrak{A}}_j$, $\sum_{a \in \mathcal{A}_j} [\overline{\pi}_t(\sigma)]_j(s,a) = \sigma_j(s)$ and $[\overline{\pi}_t(\sigma)]_j(s,a) = 0$ for all $t \in \mathcal{T}$, $\sigma \in \overline{\mathfrak{S}}$, $j \in \mathcal{J}$, $s \in \mathcal{S}_j$ and all $a \in \mathcal{A}_j \backslash \mathcal{A}_{jt}(s)$. Out of these constraints, $[\overline{\pi}_t(\sigma)]_j \in \overline{\mathfrak{A}}_j$ holds by construction, while

$$\sum_{a \in \mathcal{A}_j} [\overline{\pi}_t(\sigma)]_j(s,a) \;=\; \sum_{a \in \mathcal{A}_j} \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{\sigma^{\mathrm{LP}}_{tj}(s)} \cdot \sigma_j(s) \;=\; \sigma_j(s)$$

holds due to the fourth constraint in the fluid LP (3), and $[\overline{\pi}_t(\sigma)]_j(s,a) = 0$ is ensured by the fifth constraint of the fluid LP. We thus have $\overline{\pi} \in \overline{\Pi}$ as claimed.

We complete the proof by constructing a weakly coupled counting DP for which the fluid DP policy $\overline{\pi}$ constructed from any optimal solution $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ to the fluid LP (3) satisfies $\overline{\pi} \notin \overline{\Pi}^{\mathrm{C}}$. To this end, set $\mathcal{T} = \mathcal{J} = \{1\}$, $\mathcal{S}_1 = \{1,2\}$, $\mathcal{A}_1 = \{1\}$, $n_1 = 2$ and $q(1) = q(2) = 1/2$, while the resource constraint requires that $[\pi(\sigma)]_1(1,1) \leqslant 1$. One readily verifies that $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ defined by $\sigma^{\mathrm{LP}}_1(1) = \sigma^{\mathrm{LP}}_1(2) = 1$ and $\pi^{\mathrm{LP}}_1(1,1) = \pi^{\mathrm{LP}}_1(2,1) = 1$ is the unique feasible solution to the fluid LP (3). The corresponding fluid DP policy $\overline{\pi}$, however, necessarily violates the resource constraint in the state $\sigma_1 \in \overline{\mathfrak{S}}$ defined via $\sigma_1(1) = 2$ and $\sigma_1(2) = 0$. (Recall that a feasible policy to the fluid DP must be feasible in *every* state of the fluid DP, not just the almost sure state.) □

Proof of Proposition 3. First note that $\overline{\pi} \in \overline{\Pi}$ according to Observation 1, which implies that

$\overline{\mathfrak{A}}_t(\sigma) \neq \varnothing$ for all $t \in \mathcal{T}$ and $\sigma \in \mathfrak{S}$. From the construction of $\mathfrak{A}_t(\sigma)$ and $\overline{\mathfrak{A}}_t(\sigma)$ in the Definitions 4 and 7, respectively, one then readily verifies that $\mathfrak{A}_t(\sigma) \neq \varnothing$ for all $t \in \mathcal{T}$ and $\sigma \in \mathfrak{S}$ as well. We thus conclude that a policy $\pi^\star = \Pr(\overline{\pi})$ indeed exists and satisfies $\pi^\star \in \Pi$.

To see that $\|\pi_t^\star(\sigma) - \overline{\pi}_t(\sigma)\|_\infty < 1$ for all $t \in \mathcal{T}$ and all $\sigma \in \mathfrak{S}$, recall that

$$\pi_t^\star(\sigma) \in \operatorname*{arg\,min}_{\alpha \in \mathfrak{A}_t(\sigma)} \|\overline{\pi}_t(\sigma) - \alpha\|_1$$

by definition, where $\|\overline{\pi}_t(\sigma) - \alpha\|_1 = \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} |[\overline{\pi}_t(\sigma)]_j(s,a) - \alpha_j(s,a)|$. Since both the objective function and the constraints of this optimization problem are separable in $j \in \mathcal{J}$ and $s \in \mathcal{S}_j$, the component $\left[\Pr(\overline{\pi})_t(\sigma)\right]_j(s,\cdot)$ is any solution of the following minimization problem:

$$\begin{aligned}
\underset{\alpha}{\text{minimize}} \quad & \|[\overline{\pi}_t(\sigma)]_j(s,\cdot) - \alpha\|_1 \\
\text{subject to} \quad & \sum_{a \in \mathcal{A}_j} \alpha(a) = \sum_{a \in \mathcal{A}_j} [\overline{\pi}_t(\sigma)]_j(s,a) \\
& \alpha(a) = 0 \qquad\qquad \forall a \in \mathcal{A}_j \backslash \mathcal{A}_{jt}(s) \\
& \alpha : \mathcal{A}_j \to \mathbb{N}_0
\end{aligned}$$

We show that any solution $\alpha^\star$ to this optimization problem satisfies $\|\overline{\pi}_t(\sigma)]_j(s,\cdot) - \alpha^\star\|_\infty < 1$. Assume to the contrary that there is $a \in \mathcal{A}_j$ such that $\alpha^\star(a) \geqslant [\overline{\pi}_t(\sigma)]_j(s,a) + 1$; the proof for $\alpha^\star(a) \leqslant [\overline{\pi}_t(\sigma)]_j(s,a) - 1$ is symmetric. Since $\sum_{a \in \mathcal{A}_j} \alpha^\star(a) = \sum_{a \in \mathcal{A}_j} [\overline{\pi}_t(\sigma)]_j(s,a)$, this implies that there is another action $a' \in \mathcal{A}_{jt}(s)$ such that $\alpha^\star(a') < [\overline{\pi}_t(\sigma)]_j(s,a')$. Consider now the alternative solution α' defined via $\alpha'(a) = \alpha^\star(a) - 1$, $\alpha'(a') = \alpha^\star(a') + 1$ and $\alpha'(\cdot) = \alpha^\star(\cdot)$ elsewhere. One readily verifies that α' is also feasible, and basic algebraic manipulations show that

$$\begin{aligned}
& \left\|\alpha' - [\overline{\pi}_t(\sigma)]_j(s,\cdot)\right\|_1 \;-\; \|\alpha^\star - [\overline{\pi}_t(\sigma)]_j(s,\cdot)\|_1 \\
= \; & \left|\alpha'(a) - [\overline{\pi}_t(\sigma)]_j(s,a)\right| \;-\; |\alpha^\star(a) - [\overline{\pi}_t(\sigma)]_j(s,a)| \\
& \qquad\qquad + \; \left|\alpha'(a') - [\overline{\pi}_t(\sigma)]_j(s,a')\right| \;-\; \left|\alpha^\star(a') - [\overline{\pi}_t(\sigma)]_j(s,a')\right| \\
= \; & \big(\alpha^\star(a) - 1 - [\overline{\pi}_t(\sigma)]_j(s,a)\big) \;-\; \big(\alpha^\star(a) - [\overline{\pi}_t(\sigma)]_j(s,a)\big) \\
& \qquad\qquad + \; \left|\alpha^\star(a') + 1 - [\overline{\pi}_t(\sigma)]_j(s,a')\big)\right| \;-\; \big([\overline{\pi}_t(\sigma)]_j(s,a') - \alpha^\star(a')\big) \\
= \; & -1 \;+\; \left|\alpha^\star(a') + 1 - [\overline{\pi}_t(\sigma)]_j(s,a')\right| \;-\; \big([\overline{\pi}_t(\sigma)]_j(s,a') - \alpha^\star(a')\big) \\
= \; & -2\min\big\{[\overline{\pi}_t(\sigma)]_j(s,a') - \alpha^\star(a'),\; 1\big\} \;<\; 0,
\end{aligned}$$

where the last equality holds since the expression in the penultimate line evaluates to

$$-1 \;+\; \big([\overline{\pi}_t(\sigma)]_j(s,a') - \alpha^\star(a') - 1\big) \;-\; \big([\overline{\pi}_t(\sigma)]_j(s,a') - \alpha^\star(a')\big) \;=\; -2$$

if $\alpha^\star(a') + 1 - [\overline{\pi}_t(\sigma)]_j(s,a') \leqslant 0$ and to

$$-1 \;+\; \big(\alpha^\star(a') + 1 - [\overline{\pi}_t(\sigma)]_j(s,a')\big) \;-\; \big([\overline{\pi}_t(\sigma)]_j(s,a') - \alpha^\star(a')\big) \;=\; -2\big([\overline{\pi}_t(\sigma)]_j(s,a') - \alpha^\star(a')\big)$$

if $\alpha^\star(a') + 1 - [\overline{\pi}_t(\sigma)]_j(s,a') > 0$. This contradicts the assumed optimality of $\alpha^\star$, and we thus have $\|\overline{\pi}_t(\sigma)]_j(s,\cdot) - \alpha^\star\|_\infty < 1$ for all $j \in \mathcal{J}$, $s \in \mathcal{S}_j$. Since our arguments do not depend on the choice of $t \in \mathcal{T}$, $\sigma \in \mathfrak{S}$, we conclude that $\|\pi^\star_t(\sigma) \,-\, \overline{\pi}_t(\sigma)\|_\infty < 1$ for all $t \in \mathcal{T}$, $\sigma \in \mathfrak{S}$ as claimed. □

The proofs of Theorems 2 and 3 rely on two auxiliary results, which we prove first. In the following statements, we fix the weakly coupled counting DP $(\{\mathfrak{S}_j, \mathfrak{A}_j, \mathfrak{q}_j, \mathfrak{p}_j, \mathfrak{r}_j\}_j; \{n_j\}_j)$, a rounded policy $\pi^\star$ satisfying the conditions of Section 3.3.1 and its associated random state evolution $\tilde{\sigma}^\star = \{\tilde{\sigma}^\star_t\}_t$, as well as the quantities $\theta^\star$ and θ^{DP} from the statements of Theorems 2 and 3.

Lemma 1. *The following inequalities hold for all $t \in \mathcal{T}$, $j \in \mathcal{J}$ and $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$:*

$$\begin{aligned}
\big|\mathbb{E}\big[\tilde{\sigma}^\star_{tj}(s)\big] - \sigma^{\mathrm{LP}}_{tj}(s)\big| \;&\leqslant\; \frac{\overline{p}^t_j - 1}{\overline{p}_j - 1} - 1, \\
\big|\mathbb{E}\left[[\pi^\star_t(\tilde{\sigma}^\star_t)]_j(s,a)\right] - \pi^{\mathrm{LP}}_{tj}(s,a)\big| \;&\leqslant\; \frac{\overline{p}^t_j - 1}{\overline{p}_j - 1}.
\end{aligned}$$

Proof of Lemma 1. We prove the statement via induction over $t \in \mathcal{T}$. For $t = 1$, Definition 4 of a weakly coupled counting DP states that

$$\mathbb{E}\big[\tilde{\sigma}^\star_{1j}(s)\big] \;=\; n_j \cdot q_j(s) \;=\; \sigma^{LP}_{1j}(s) \qquad \forall j \in \mathcal{J},\ \forall s \in \mathcal{S}_j, \tag{4}$$

and Proposition 3 implies that

$$\left|\mathbb{E}\left[[\pi^\star_1(\tilde{\sigma}^\star_1)]_j(s,a)\right] - \mathbb{E}\left[[\overline{\pi}_1(\tilde{\sigma}^\star_1)]_j(s,a)\right]\right| \;\leqslant\; 1 \qquad \forall j \in \mathcal{J},\ \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j.$$

Moreover, we have

$$\mathbb{E}\left[[\overline{\pi}_1(\tilde{\sigma}_1^\star)]_j(s,a)\right] \;=\; \frac{\pi_{1j}^{\mathrm{LP}}(s,a)}{\sigma_{1j}^{\mathrm{LP}}(s)} \cdot \mathbb{E}\left[\tilde{\sigma}_{1j}^\star(s)\right] \;=\; \pi_{1j}^{\mathrm{LP}}(s) \qquad \forall j \in \mathcal{J},\ \forall s \in \mathcal{S}_j,$$

where the first identity follows from the definition of $\overline{\pi}$, while the second identity is due to (4). This proves the statement for $t = 1$.

Assume now that the statement holds for some $t \in \mathcal{T}\backslash\{T\}$. We then have

$$\mathbb{E}\left[\tilde{\sigma}_{t+1,j}^\star(s') \,|\, \pi_t^\star(\tilde{\sigma}_t^\star)\right] \;=\; \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a) \cdot [\pi_t^\star(\tilde{\sigma}_t^\star)]_j(s,a) \qquad \forall j \in \mathcal{J},\ \forall s' \in \mathcal{S}_j.$$

Taking expectation on both sides, we see that for all $j \in \mathcal{J}$ and all $s' \in \mathcal{S}_j$, we have that

$$\begin{aligned}
\mathbb{E}\left[\tilde{\sigma}_{t+1,j}^\star(s')\right] \;&=\; \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a) \cdot \mathbb{E}\left[[\pi_t^\star(\tilde{\sigma}_t^\star)]_j(s,a)\right] \\
&=\; \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a) \cdot \left(\pi_{tj}^{\mathrm{LP}}(s,a) + \mathbb{E}\left[[\pi_t^\star(\tilde{\sigma}_t^\star)]_j(s,a)\right] - \pi_{tj}^{\mathrm{LP}}(s,a)\right) \\
&=\; \sigma_{t+1,j}^{\mathrm{LP}}(s') + \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a) \left(\mathbb{E}\left[[\pi_t^\star(\tilde{\sigma}_t^\star)]_j(s,a)\right] - \pi_{tj}^{\mathrm{LP}}(s,a)\right),
\end{aligned}$$

where the last identity follows from the second constraint of the fluid LP (3), which implies that $\sigma_{t+1,j}^{\mathrm{LP}}(s') = \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a) \cdot \pi_{tj}^{\mathrm{LP}}(s,a)$ for all $t \in \mathcal{T}\backslash\{T\}$, $j \in \mathcal{J}$ and $s' \in \mathcal{S}_j$. Subtracting $\sigma_{t+1,j}^{\mathrm{LP}}(s')$ on both sides and taking absolute values, we see that

$$\begin{aligned}
\left|\mathbb{E}\left[\tilde{\sigma}_{t+1,j}^\star(s')\right] - \sigma_{t+1,j}^{\mathrm{LP}}(s')\right| \;&=\; \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a) \left|\mathbb{E}\left[[\pi_t^\star(\tilde{\sigma}_t^\star)]_j(s,a)\right] - \pi_{tj}^{\mathrm{LP}}(s,a)\right| \\
&\leqslant\; \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a) \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1} \\
&\leqslant\; \overline{p}_j \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1} \;=\; \frac{\overline{p}_j^{t+1} - 1}{\overline{p}_j - 1} - 1,
\end{aligned}$$

where the two inequalities follow from the induction hypothesis and the definition of $\overline{p}_j$, respectively.

Similarly, the definition of $\overline{\pi}$ implies that

$$\begin{aligned}
\left|\mathbb{E}\left[[\overline{\pi}_{t+1}(\tilde{\sigma}^{\star}_{t+1})]_j(s,a)\right] - \pi^{\mathrm{LP}}_{t+1,j}(s,a)\right| &= \left|\frac{\pi^{\mathrm{LP}}_{t+1,j}(s,a)}{\sigma^{\mathrm{LP}}_{t+1,j}(s)} \cdot \mathbb{E}\left[\tilde{\sigma}^{\star}_{t+1,j}(s)\right] - \pi^{\mathrm{LP}}_{t+1,j}(s,a)\right| \\
&= \frac{\pi^{\mathrm{LP}}_{t+1,j}(s,a)}{\sigma^{\mathrm{LP}}_{t+1,j}(s)} \cdot \left|\mathbb{E}\left[\tilde{\sigma}^{\star}_{t+1,j}(s)\right] - \sigma^{\mathrm{LP}}_{t+1,j}(s)\right| \\
&\leqslant \left|\mathbb{E}\left[\tilde{\sigma}^{\star}_{t+1,j}(s)\right] - \sigma^{\mathrm{LP}}_{t+1,j}(s)\right| \\
&\leqslant \frac{\overline{p}_j^{t+1} - 1}{\overline{p}_j - 1} - 1,
\end{aligned} \tag{5}$$

where the first identity holds by definition of $\overline{\pi}_{t+1}$, the first inequality follows from the fact that $\pi^{\mathrm{LP}}_{t+1,j}(s,a) \leqslant \sigma^{\mathrm{LP}}_{t+1,j}(s)$, which is implied by the fourth constraint of the fluid LP (3), and the second inequality is due to the induction hypothesis. Proposition 3 implies that

$$\left|\mathbb{E}\left[[\pi^{\star}_{t+1}(\tilde{\sigma}^{\star}_{t+1})]_j(s,a)\right] - \mathbb{E}\left[[\overline{\pi}_{t+1}(\tilde{\sigma}^{\star}_{t+1})]_j(s,a)\right]\right| \leqslant 1 \qquad \forall j \in \mathcal{J},\ \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j,$$

which in turn implies through the triangle inequality that

$$\left|\mathbb{E}\left[[\pi^{\star}_{t+1}(\tilde{\sigma}^{\star}_{t+1})]_j(s,a)\right] - \pi^{\mathrm{LP}}_{t+1,j}(s,a)\right| \leqslant 1 + \left|\mathbb{E}\left[[\overline{\pi}_{t+1}(\tilde{\sigma}^{\star}_{t+1})]_j(s,a)\right] - \pi^{\mathrm{LP}}_{t+1,j}(s,a)\right|$$

for all $j \in \mathcal{J}$ and all $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$, and equation (5) implies that the right-hand side expression is less than or equal to $(\overline{p}_j^{t+1} - 1)/(\overline{p}_j - 1)$ as desired. This completes the proof. □

Lemma 2. *With probability at least* $1 - 2|\mathcal{T}| \cdot \sum_{j \in \mathcal{J}} |\mathcal{S}_j|/n_j$*, we have*

$$\left|[\pi^{\star}_t(\tilde{\sigma}^{\star}_t)]_j(s,a) - \pi^{\mathrm{LP}}_{tj}(s,a)\right| \leqslant (1 + \epsilon n_j) \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1} \qquad \forall t \in \mathcal{T},\ \forall j \in \mathcal{J},\ \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j.$$

Proof of Lemma 2. We prove the statement via induction over $t \in \mathcal{T}$. For $t = 1$, Proposition 3 implies that

$$\left|[\pi^{\star}_1(\tilde{\sigma}^{\star}_1)]_j(s,a) - [\overline{\pi}_1(\tilde{\sigma}^{\star}_1)]_j(s,a)\right| \leqslant 1 \quad \forall j \in \mathcal{J},\ \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j$$

pointwise. Moreover, we have

$$[\overline{\pi}_1(\tilde{\sigma}_1^\star)]_j(s,a) \;=\; \frac{\pi_{1j}^{\mathrm{LP}}(s,a)}{\sigma_{1j}^{\mathrm{LP}}(s)} \cdot \tilde{\sigma}_{1j}^\star(s) \;=\; \pi_{1j}^{\mathrm{LP}}(s,a) + \pi_{1j}^{\mathrm{LP}}(s,a) \cdot \frac{\tilde{\sigma}_{1j}^\star(s) - \sigma_{1j}^{\mathrm{LP}}(s)}{\sigma_{1j}^{\mathrm{LP}}(s)}$$

for all $j \in \mathcal{J}$ and $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$ pointwise, where the first identity follows from the definition of $\overline{\pi}$. This in turn implies that

$$\left|[\overline{\pi}_1(\tilde{\sigma}_1^\star)]_j(s,a) - \pi_{1j}^{\mathrm{LP}}(s,a)\right| \;=\; \left|\pi_{1j}^{\mathrm{LP}}(s,a) \cdot \frac{\tilde{\sigma}_{1j}^\star(s) - \sigma_{1j}^{\mathrm{LP}}(s)}{\sigma_{1j}^{\mathrm{LP}}(s)}\right| \;\leqslant\; \left|\tilde{\sigma}_{1j}^\star(s) - \sigma_{1j}^{\mathrm{LP}}(s)\right|$$

for all $j \in \mathcal{J}$ and $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$ pointwise, where the inequality follows from the fact that $\pi_{1j}^{\mathrm{LP}}(s,a) \,/\, \sigma_{1j}^{\mathrm{LP}}(s) \leqslant 1$ due to the fourth constraint in the fluid LP (3) (also for $\sigma_{1j}^{\mathrm{LP}}(s) = 0$, in which case our earlier convention implies that the fraction vanishes). The statement for $t = 1$ now follows from the triangle inequality if we can show that $\left|\tilde{\sigma}_{1j}^\star(s) - \sigma_{1j}^{\mathrm{LP}}(s)\right| \leqslant 1 + \epsilon n_j$ simultaneously for all $j \in \mathcal{J}$ and $s \in \mathcal{S}_j$ with probability at least $1 - 2|\mathcal{T}| \cdot \sum_{j \in \mathcal{J}} |\mathcal{S}_j|/n_j$. To see this, we note that

$$\mathbb{P}\left[\left|\tilde{\sigma}_{1j}^\star(s) - n_j \cdot q_j(s)\right| \;\leqslant\; \epsilon n_j\right] \;\geqslant\; 1 - \frac{2}{n_j} \qquad \forall j \in \mathcal{J},\; \forall s \in \mathcal{S}_j$$

according to Hoeffding's inequality, and $\sigma_{1j}^{\mathrm{LP}}(s) = n_j \cdot q_j(s)$ according to the first constraint in the fluid LP (3). The result then follows from the union bound.

Assume now that the statement holds for $t \in \mathcal{T} \backslash \{T\}$. The same argument as before shows that

$$\left|[\pi_{t+1}^\star(\tilde{\sigma}_{t+1}^\star)]_j(s,a) - [\overline{\pi}_{t+1}(\tilde{\sigma}_{t+1}^\star)]_j(s,a)\right| \;\leqslant\; 1 \tag{6}$$

as well as

$$\left|[\overline{\pi}_{t+1}(\tilde{\sigma}_{t+1}^\star)]_j(s,a) - \pi_{t+1,j}^{\mathrm{LP}}(s,a)\right| \;\leqslant\; \left|\tilde{\sigma}_{t+1,j}^\star(s) - \sigma_{t+1,j}^{\mathrm{LP}}(s)\right|$$

for all $j \in \mathcal{J}$ and $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$ pointwise. The result again follows from the triangle inequality if we can show that

$$\left|\tilde{\sigma}_{t+1,j}^\star(s) - \sigma_{t+1,j}^{\mathrm{LP}}(s)\right| \leqslant (1 + \epsilon n_j) \cdot \frac{\overline{p}_j^{t+1} - 1}{\overline{p}_j - 1} - 1 \tag{7}$$

simultaneously for all $j \in \mathcal{J}$ and $s \in \mathcal{S}_j$ with probability at least $1 - 2|\mathcal{T}| \cdot \sum_{j \in \mathcal{J}} |\mathcal{S}_j|/n_j$. The remainder of the proof is thus dedicated to proving the bound (7).

Note first that

$$\mathbb{E}\left[\tilde{\sigma}^{\star}_{t+1,j}(s') \,|\, \pi^{\star}_t(\tilde{\sigma}^{\star}_t)\right] \;=\; \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a)\cdot[\pi^{\star}_t(\tilde{\sigma}^{\star}_t)]_j(s,a) \qquad \forall j\in\mathcal{J},\ \forall s'\in\mathcal{S}_j.$$

Hoeffding's inequality, which applies since the random variables $\tilde{\sigma}^{\star}_{t+1,j}(s')$ are conditionally independent given $\pi^{\star}_t(\tilde{\sigma}^{\star}_t)$, then implies that

$$\mathbb{P}\left[\left|\tilde{\sigma}^{\star}_{t+1,j}(s') - \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s|s,a)\cdot\pi^0_{tj}(s,a)\right| \leqslant \epsilon n_j \;\middle|\; \pi^{\star}_t(\tilde{\sigma}^{\star}_t) \;=\; \pi^0_t\right] \geqslant 1-2\mathrm{e}^{-2\epsilon^2 n_j} \geqslant 1-\frac{2}{n_j}$$

for all $j\in\mathcal{J}$ and $s'\in\mathcal{S}_j$, and an application of the union bound shows that

$$\mathbb{P}\left[\left|\tilde{\sigma}^{\star}_{t+1,j}(s') - \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a)\cdot\pi^0_{tj}(s,a)\right| \leqslant \epsilon n_j \;\;\forall j\in\mathcal{J},\ \forall s'\in\mathcal{S}_j \;\middle|\; \pi^{\star}_t(\tilde{\sigma}^{\star}_t) \;=\; \pi^0_t\right] \geqslant 1-\sum_{j\in\mathcal{J}}\frac{2|\mathcal{S}_j|}{n_j}. \tag{8}$$

Note next that for any $\pi^0_t\in\mathfrak{A}$, we have

$$\begin{aligned}
&\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a)\cdot\pi^0_{tj}(s,a)\\
&= \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a)\cdot\left(\pi^0_{tj}(s,a)-\pi^{\mathrm{LP}}_{tj}(s,a)\right) \;+\; \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a)\cdot\pi^{\mathrm{LP}}_{tj}(s,a)\\
&= \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a)\cdot\left(\pi^0_{tj}(s,a)-\pi^{\mathrm{LP}}_{tj}(s,a)\right) \;+\; \sigma^{\mathrm{LP}}_{t+1,j}(s')
\end{aligned} \tag{9}$$

for all $s'\in\mathcal{S}_j$, where the last identity follows from the second constraint of the fluid LP (3). Consider next the set

$$\Gamma \;=\; \left\{\pi_t\in\mathfrak{A} \,:\, \left|\pi_{tj}(s,a)-\pi^{\mathrm{LP}}_{tj}(s,a)\right| \leqslant (1+\epsilon n_j)\cdot\frac{\overline{p}^t_j-1}{\overline{p}_j-1} \;\;\forall j\in\mathcal{J},\ \forall(s,a)\in\mathcal{S}_j\times\mathcal{A}_j\right\}$$

of partial policies π_t that are sufficiently close to π^{LP}_t. For any $\pi^0_t\in\Gamma$, we have

$$\mathbb{P}\left[\left|\tilde{\sigma}^{\star}_{t+1,j}(s') - \sigma^{\mathrm{LP}}_{t+1,j}(s')\right| \leqslant (1+\epsilon n_j)\cdot\frac{\overline{p}^{t+1}_j-1}{\overline{p}_j-1}-1 \;\;\forall j\in\mathcal{J},\ \forall s'\in\mathcal{S}_j \;\middle|\; \pi^{\star}_t(\tilde{\sigma}^{\star}_t)=\pi^0_t\right] \geqslant 1-\sum_{j\in\mathcal{J}}\frac{2|\mathcal{S}_j|}{n_j} \tag{10}$$

since

$$
\begin{aligned}
& \left|\tilde{\sigma}^{\star}_{t+1,j}(s') - \sigma^{\mathrm{LP}}_{t+1,j}(s')\right| \\
\leqslant\ & \left|\tilde{\sigma}^{\star}_{t+1,j}(s') - \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a)\cdot\pi^0_{tj}(s,a)\right| \ +\ \left|\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a)\cdot\pi^0_{tj}(s,a) - \sigma^{\mathrm{LP}}_{t+1,j}(s')\right| \\
\leqslant\ & \epsilon n_j \ +\ \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a)\cdot(1+\epsilon n_j)\cdot\frac{\overline{p}_j^t - 1}{\overline{p}_j - 1} \\
\leqslant\ & \epsilon n_j \ +\ (1+\epsilon n_j)\cdot\frac{\overline{p}_j^{t+1} - \overline{p}_j}{\overline{p}_j - 1} \quad = \quad (1+\epsilon n_j)\cdot\frac{\overline{p}_j^{t+1} - 1}{\overline{p}_j - 1} - 1,
\end{aligned}
$$

where the first inequality holds pointwise due to the triangle inequality, the second inequality holds conditionally with probability at least $1 - 2\sum_{j\in\mathcal{J}}|\mathcal{S}_j|/n_j$ due to (8), (9) and the fact that $\pi^0_t \in \Gamma$, and the third inequality holds by definition of $\overline{p}_j$.

Going over to unconditional probabilities, we finally obtain

$$
\begin{aligned}
& \mathbb{P}\left[\left|\tilde{\sigma}^{\star}_{t+1,j}(s') - \sigma^{\mathrm{LP}}_{t+1,j}(s')\right| \leqslant (1+\epsilon n_j)\cdot\frac{\overline{p}_j^{t+1} - 1}{\overline{p}_j - 1} - 1 \ \ \forall j\in\mathcal{J},\ \forall s'\in\mathcal{S}_j\right] \\
=\ & \sum_{\pi^0_t}\mathbb{P}\left[\left|\tilde{\sigma}^{\star}_{t+1,j}(s') - \sigma^{\mathrm{LP}}_{t+1,j}(s')\right| \leqslant (1+\epsilon n_j)\cdot\frac{\overline{p}_j^{t+1} - 1}{\overline{p}_j - 1} - 1 \ \ \forall j\in\mathcal{J},\ \forall s'\in\mathcal{S}_j \ \middle|\ \pi^{\star}_t(\tilde{\sigma}^{\star}_t) = \pi^0_t\right] \\
& \qquad\qquad \cdot\,\mathbb{P}\left[\pi^{\star}_t(\tilde{\sigma}^{\star}_t) = \pi^0_t\right] \\
\geqslant\ & \sum_{\pi^0_t\in\Gamma}\mathbb{P}\left[\left|\tilde{\sigma}^{\star}_{t+1,j}(s') - \sigma^{\mathrm{LP}}_{t+1,j}(s')\right| \leqslant (1+\epsilon n_j)\cdot\frac{\overline{p}_j^{t+1} - 1}{\overline{p}_j - 1} - 1 \ \ \forall j\in\mathcal{J},\ \forall s'\in\mathcal{S}_j \ \middle|\ \pi^{\star}_t(\tilde{\sigma}^{\star}_t) = \pi^0_t\right] \\
& \qquad\qquad \cdot\,\mathbb{P}\left[\pi^{\star}_t(\tilde{\sigma}^{\star}_t) = \pi^0_t\right] \\
\geqslant\ & \sum_{\pi^0_t\in\Gamma}\left(1 - \sum_{j\in\mathcal{J}}\frac{2|\mathcal{S}_j|}{n_j}\right)\cdot\mathbb{P}\left[\pi^{\star}_t(\tilde{\sigma}^{\star}_t) = \pi^0_t\right] \\
\geqslant\ & \left(1 - \sum_{j\in\mathcal{J}}\frac{2|\mathcal{S}_j|}{n_j}\right)\cdot\left(1 - t\cdot\sum_{j\in\mathcal{J}}\frac{2|\mathcal{S}_j|}{n_j}\right) \ \geqslant\ 1 - (t+1)\cdot\sum_{j\in\mathcal{J}}\frac{2|\mathcal{S}_j|}{n_j},
\end{aligned}
$$

where the identity is due to the law of total probability, the first inequality holds because we restrict ourselves to $\pi^0_t \in \Gamma$, the second inequality follows from (10), the third inequality is due to the definition of Γ as well as the induction hypothesis, and the last inequality holds since $(1-x)(1-tx) = 1 - (t+1)x + tx^2 \geqslant 1 - (t+1)x$ for all $x \in \mathbb{R}$. This shows the bound (7) and thereby completes the proof. □

Proof of Theorem 2. In view of the bound on the expected total reward, we observe that

$$\begin{aligned}
\theta^\star &= \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\mathbb{E}\left[[\pi_t^\star(\tilde{\sigma}_t^\star)]_j(s,a)\right] \\
&\geqslant \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\left(\pi_{tj}^{\mathrm{LP}}(s,a)-\frac{\overline{p}_j^t-1}{\overline{p}_j-1}\right) \\
&= \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\pi_{tj}^{\mathrm{LP}}(s,a) \;-\; \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\frac{\overline{p}_j^t-1}{\overline{p}_j-1} \\
&= \theta^{\mathrm{LP}} \;-\; \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\frac{\overline{p}_j^t-1}{\overline{p}_j-1} \\
&\geqslant \theta^{\mathrm{DP}} \;-\; \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\frac{\overline{p}_j^t-1}{\overline{p}_j-1},
\end{aligned}$$

where the first identity holds by definition of $\theta^\star$, the first inequality follows from Lemma 1, the third identity holds by definition of θ^{LP}, and the second inequality is due to Theorem 1.

As for the resource violation, we have with probability at least $1-2|\mathcal{T}|\cdot\sum_{j\in\mathcal{J}}|\mathcal{S}_j|/n_j$ that

$$\begin{aligned}
&\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot[\pi_t^\star(\tilde{\sigma}_t^\star)]_j(s,a) \\
\leqslant\; &\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot\left(\pi_{tj}^{\mathrm{LP}}(s,a)+(1+\epsilon n_j)\cdot\frac{\overline{p}_j^t-1}{\overline{p}_j-1}\right) \\
=\; &\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot\pi_{tj}^{\mathrm{LP}}(s,a)+\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j}(1+\epsilon n_j)\cdot\frac{\overline{p}_j^t-1}{\overline{p}_j-1}\cdot c_{tlj}(s,a) \\
\leqslant\; &b_{tl}+\sum_{j\in\mathcal{J}}(1+\epsilon n_j)\cdot\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot\frac{\overline{p}_j^t-1}{\overline{p}_j-1},
\end{aligned}$$

where the first inequality is due to Lemma 2, and the second inequality holds since π^{LP} is a feasible solution to the fluid LP (3) and hence satisfies the third constraint of the LP. □

Proof of Theorem 3. The bound on the resource violation is the same as in Theorem 2, and we refer to its proof for the justification of the bound. In view of the bound on the total reward,

we observe that with probability at least $1 - 2|\mathcal{T}| \cdot \sum_{j\in\mathcal{J}} |\mathcal{S}_j|/n_j$, we have

$$\begin{aligned}
\tilde{\theta}^\star &= \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} r_{jt}(s,a) \cdot [\pi_t^\star(\tilde{\sigma}_t^\star)]_j(s,a) \\
&\geqslant \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} r_{jt}(s,a) \cdot \left(\pi_{tj}^{\mathrm{LP}}(s,a) - (1+\epsilon n_j) \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1} \right) \\
&= \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} r_{jt}(s,a) \cdot \pi_{tj}^{\mathrm{LP}}(s,a) \; - \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} r_{jt}(s,a) \cdot (1+\epsilon n_j) \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1} \\
&= \theta^{\mathrm{LP}} \; - \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} r_{jt}(s,a) \cdot (1+\epsilon n_j) \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1} \\
&\geqslant \theta^{\mathrm{DP}} \; - \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} r_{jt}(s,a) \cdot (1+\epsilon n_j) \cdot \frac{\overline{p}_j^t - 1}{\overline{p}_j - 1},
\end{aligned}$$

where the first identity holds pointwise by definition of $\tilde{\theta}^\star$, the first inequality is due to Lemma 2, the third identity holds by definition of θ^{LP}, and the second inequality is due to Theorem 1. □

Proof of Proposition 4. Fix any $t \in \mathcal{T}$, $\sigma \in \mathfrak{S}$, $j \in \mathcal{J}$ and $s \in \mathcal{S}_j$. By construction, $[\pi_t^\star(\sigma)]_j(s,\cdot)$ is any solution to the problem

$$\begin{aligned}
\underset{\alpha}{\text{minimize}} \quad & \left\| [\overline{\pi}_t(\sigma)]_j(s,\cdot) - \alpha \right\|_1 \\
\text{subject to} \quad & \sum_{a\in\mathcal{A}_j} \alpha(a) = \sum_{a\in\mathcal{A}_j} [\overline{\pi}_t(\sigma)]_j(s,a) \\
& \alpha(a) = 0 \qquad\qquad \forall a \in \mathcal{A}_j \backslash \mathcal{A}_{jt}(s) \\
& \alpha : \mathcal{A}_j \to \mathbb{N}_0.
\end{aligned}$$

By Proposition 3, we have $\left\| \pi_t^\star(\sigma) - \overline{\pi}_t(\sigma) \right\|_\infty < 1$, which implies that

$$\lfloor [\overline{\pi}_t(\sigma)]_j(s,a) \rfloor \;\leqslant\; [\pi_t^\star(\sigma)]_j(s,a) \;\leqslant\; \lfloor [\overline{\pi}_t(\sigma)]_j(s,a) \rfloor + 1 \qquad \forall a \in \mathcal{A}_j.$$

The substitutions $\beta \leftarrow \mathrm{frac}([\overline{\pi}_t(\sigma)]_j(s,\cdot))$ and $\alpha \leftarrow \alpha - \lfloor [\overline{\pi}_t(\sigma)]_j(s,\cdot) \rfloor$ thus imply that $[\pi_t^\star(\sigma)]_j(s,\cdot) -$

$\lfloor [\overline{\pi}_t(\sigma)]_j(s,\cdot) \rfloor$ is any solution to the problem

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & \|\beta - \alpha\|_1 \\ \text{subject to} \quad & \sum_{a \in \mathcal{A}_j} \alpha(a) = \sum_{a \in \mathcal{A}_j} \beta(\alpha) \\ & \alpha(a) = 0 \qquad\qquad \forall a \in \mathcal{A}_j \backslash \mathcal{A}_{jt}(s) \\ & \alpha : \mathcal{A}_j \to \{0,1\}. \end{aligned}$$

Define the set $\mathcal{A}_j(s,\alpha) = \{a \in \mathcal{A}_j \,:\, \alpha(a) = 1\}$ and observe that $|\mathcal{A}_j(s,\alpha)| = \sum_{a \in \mathcal{A}_j} \alpha(a) = \sum_{a \in \mathcal{A}_j} \beta(a) = \|\beta\|_1$. Moreover, the objective function in the above problem evaluates to

$$\begin{aligned} \|\beta - \alpha\|_1 \;&=\; \sum_{a \in \mathcal{A}_j(s,\alpha)} (1 - \beta(a)) \;+\; \sum_{a \notin \mathcal{A}_j(s,\alpha)} \beta(a) \\ &=\; |\mathcal{A}_j(s,\alpha)| \;-\; 2 \sum_{a \in \mathcal{A}_j(s,\alpha)} \beta(a) \;+\; \sum_{a \in \mathcal{A}_j} \beta(a) \;=\; 2\,\|\beta\|_1 \;-\; 2 \sum_{a \in \mathcal{A}_j(s,\alpha)} \beta(a). \end{aligned}$$

Since $|\mathcal{A}_j(s,\alpha)| = \|\beta\|_1$ does not depend on the choice of α (as long as α is feasible), the optimal choice of $\mathcal{A}_j(s,\alpha)$ consists of the $\|\beta\|_1$ largest entries of β. Moreover, any optimal policy $\pi^\star$ must satisfy $\left[\pi_t^\star(\sigma)\right]_j(s,a) = \lfloor [\overline{\pi}_t(\sigma)]_j(s,a) \rfloor + \mathbf{1}[a \in \mathcal{A}_j(s,\alpha)]$, $a \in \mathcal{A}_j$. The statement then follows from the fact that $\mathcal{I}_{tj}(\sigma) = \{(s,a) \in \mathcal{S}_j \times \mathcal{A}_j \,:\, a \in \mathcal{A}_j(s,\alpha)\}$. □

The proofs of Theorems 4 and 5 rely on the following two auxiliary results, which we prove first.

Lemma 3. *The following equations hold for all $t \in \mathcal{T}$, $(j,i) \in \mathcal{J} \times \{1,\ldots,n_j\}$ and $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$:*

$$\mathbb{P}\left[\tilde{s}_{t,(j,i)} = s \,\wedge\, \tilde{a}_{t,(j,i)} = a\right] \;=\; \frac{\pi_{tj}^{\mathrm{LP}}(s,a)}{n_j}.$$

Proof of Lemma 3. According to our definition of the randomized policy $\pi^\star$, we have

$$\mathbb{P}\left[\tilde{a}_{t,(j,i)} = a \,|\, \tilde{s}_{t,(j,i)} = s\right] \;=\; \frac{\pi_{tj}^{\mathrm{LP}}(s,a)}{\sigma_{tj}^{\mathrm{LP}}(s)} \qquad \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j$$

for each $t \in \mathcal{T}$ and $(j,i) \in \mathcal{J} \times \{1,\ldots,n_j\}$, which in turn implies that

$$\mathbb{P}\left[\tilde{s}_{t,(j,i)} = s \,\wedge\, \tilde{a}_{t,(j,i)} = a\right] \;=\; \mathbb{P}\left[\tilde{a}_{t,(j,i)} = a \,|\, \tilde{s}_{t,(j,i)} = s\right] \cdot \mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right] = \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{\sigma^{\mathrm{LP}}_{tj}(s)} \cdot \mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right]. \tag{11}$$

In the remainder of the proof, we show via induction on $t \in \mathcal{T}$ that $\mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right] = \sigma^{\mathrm{LP}}_{tj}(s)\,/\,n_j$ for all $t \in \mathcal{T}$, $(j,i) \in \mathcal{J} \times \{1,\ldots,n_j\}$ and $s \in \mathcal{S}_j$, which concludes the proof.

For $t = 1$, the definition of the weakly coupled DP implies that $\mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right] = q_j(s)$, and the first constraint of the fluid LP (3) ensures that $\sigma^{\mathrm{LP}}_{1j}(s) = n_j \cdot q_j(s)$. Assume now that $\mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right] = \sigma^{\mathrm{LP}}_{tj}(s)\,/\,n_j$ for some $t \in \mathcal{T}\backslash\{T\}$. We then have

$$\begin{aligned}
\mathbb{P}\left[\tilde{s}_{t+1,(j,i)} = s'\right] \;&=\; \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a) \cdot \mathbb{P}\left[\tilde{s}_{t,(j,i)} = s \,\wedge\, \tilde{a}_{t,(j,i)} = a\right] \\
&=\; \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a) \cdot \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{\sigma^{\mathrm{LP}}_{tj}(s)} \cdot \mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right] \\
&=\; \sum_{s\in\mathcal{S}_j} \sum_{a\in\mathcal{A}_j} p_{jt}(s'|s,a) \cdot \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{n_j} \;=\; \frac{\sigma^{\mathrm{LP}}_{t+1,j}(s')}{n_j},
\end{aligned}$$

where the first identity follows from the definition of the weakly coupled DP, the second identity is due to (11), the third identity follows from the induction hypothesis, and the last identity is due to the second constraint of the fluid LP (3). □

Lemma 4. *Let $\tilde{\alpha}^{\star}_{tj}(s,a) = \sum_{i=1}^{n_j} \mathbf{1}\left[\tilde{s}_{t,(j,i)} = s \,\wedge\, \tilde{a}_{t,(j,i)} = a\right]$ record the number of DPs in the j-th counting DP that are in state s and to which action a is applied at time t. Then*

$$\mathbb{E}\left[\tilde{\alpha}^{\star}_{tj}(s,a)\right] \;=\; \pi^{\mathrm{LP}}_{tj}(s,a) \qquad \forall t \in \mathcal{T},\; \forall j \in \mathcal{J},\; \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j.$$

Furthermore, with probability at least $1 - 2|\mathcal{T}| \cdot \sum_{j\in\mathcal{J}} |\mathcal{S}_j| \cdot |\mathcal{A}_j|\,/\,n_j$, we have

$$\left|\tilde{\alpha}^{\star}_{tj}(s,a) - \pi^{\mathrm{LP}}_{tj}(s,a)\right| \;\leqslant\; \epsilon n_j \qquad \forall t \in \mathcal{T},\; \forall j \in \mathcal{J},\; \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j.$$

Proof of Lemma 4. In view of the first statement, we note that

$$\mathbb{E}\left[\tilde{\alpha}^{\star}_{tj}(s,a)\right] \;=\; \sum_{i=1}^{n_j} \mathbb{P}\left[\tilde{s}_{t,(j,i)} = s \,\wedge\, \tilde{a}_{t,(j,i)} = a\right] \;=\; \sum_{i=1}^{n_j} \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{n_j} \;=\; \pi^{\mathrm{LP}}_{tj}(s,a),$$

where the first and second identity follow from the definition of $\tilde{\alpha}^{\star}_{tj}(s,a)$ and Lemma 3, respectively.

As for the second statement, note that $\tilde{\alpha}^{\star}_{tj}(s,a)$ is a sum of i.i.d. random variables. Hoeffding's inequality then implies that

$$\mathbb{P}\left[\left|\tilde{\alpha}^{\star}_{tj}(s,a) - \pi^{\mathrm{LP}}_{tj}(s,a)\right| \leqslant \epsilon n_j\right] \;\geqslant\; 1 - 2\mathrm{e}^{2\epsilon^2 n_j} \;\geqslant\; 1 - \frac{2}{n_j}$$

for all $t \in \mathcal{T}$, $j \in \mathcal{J}$ and $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$, and the statement thus follows from the union bound. $\square$

Proof of Theorem 4. In view of the bound on the expected total reward, we have

$$\begin{aligned}
\theta^{\star} \;&=\; \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{i=1}^{n_j} r_{jt}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)})\right] \\
&=\; \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{i=1}^{n_j}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot \mathbf{1}\left[\tilde{s}_{t,(j,i)} = s \;\wedge\; \tilde{a}_{t,(j,i)} = a\right]\right] \\
&=\; \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot \mathbb{E}\left[\tilde{\alpha}^{\star}_{tj}(s,a)\right] \\
&=\; \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot \pi^{\mathrm{LP}}_{tj}(s,a) \\
&=\; \theta^{\mathrm{LP}} \;\geqslant\; \theta^{\mathrm{DP}},
\end{aligned}$$

where the first identity holds by definition of $\theta^{\star}$, the third identity follows from the definition of $\alpha^{\star}_{tj}(s,a)$ in Lemma 4, the fourth identity is due to Lemma 4, the last identity holds by definition of θ^{LP}, and the inequality follows from Theorem 1.

As for the resource violation, with probability at least $1 - 2|\mathcal{T}| \cdot \sum_{j\in\mathcal{J}} |\mathcal{S}_j| \cdot |\mathcal{A}_j| \,/\, n_j$ we have

$$\begin{aligned}
\sum_{j\in\mathcal{J}}\sum_{i=1}^{n_j} c_{tlj}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)}) \;&=\; \sum_{j\in\mathcal{J}}\sum_{i=1}^{n_j}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot \mathbf{1}\left[\tilde{s}_{t,(j,i)} = s \;\wedge\; \tilde{a}_{t,(j,i)} = a\right] \\
&=\; \sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot \tilde{\alpha}^{\star}_{tj}(s,a) \\
&\leqslant\; \sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot \left(\pi^{\mathrm{LP}}_{tj}(s,a) + \epsilon n_j\right) \\
&\leqslant\; b_{tl} + \epsilon \sum_{j\in\mathcal{J}} n_j \cdot \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)
\end{aligned}$$

for all $l \in \mathcal{L}$ and $t \in \mathcal{T}$, where the second identity follows from the definition of $\tilde{\alpha}^{\star}_{tj}(s,a)$ in Lemma 4, the first inequality is due to the statement of Lemma 4, and the second inequality is implied by the third constraint of the fluid LP (3). $\square$

Proof of Theorem 5. The bound on the resource violation is the same as in Theorem 4, and we refer to its proof for the justification of the bound. In view of the bound on the expected total reward, we observe that with probability at least $1 - 2|\mathcal{T}| \cdot \sum_{j \in \mathcal{J}} |\mathcal{S}_j| \cdot |\mathcal{A}_j| \,/\, n_j$, we have

$$
\begin{aligned}
\tilde{\theta}^{\star} &= \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} r_{jt}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)}) \\
&= \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} r_{jt}(s,a) \cdot \tilde{\alpha}^{\star}_{tj}(s,a) \\
&\geqslant \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} r_{jt}(s,a) \cdot \big(\pi^{\mathrm{LP}}_{tj}(s,a) - \epsilon n_j\big) \\
&= \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} r_{jt}(s,a) \cdot \pi^{\mathrm{LP}}_{tj}(s,a) \;-\; \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} r_{jt}(s,a) \cdot \epsilon n_j \\
&= \theta^{\mathrm{LP}} \;-\; \epsilon \cdot \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} n_j \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} r_{jt}(s,a) \\
&\geqslant \theta^{\mathrm{DP}} \;-\; \epsilon \cdot \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} n_j \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} r_{jt}(s,a),
\end{aligned}
$$

where the first identity is due to the definition of $\tilde{\theta}^{\star}$, the second identity follows from the definition of $\tilde{\alpha}^{\star}_{tj}(s,a)$ in Lemma 4, the first inequality is due to the statement of Lemma 4, the last identity follows from the definition of θ^{LP}, and the second inequality is due to Theorem 1. $\square$