DYNAMICALLY OPTIMAL TREATMENT ALLOCATION USING REINFORCEMENT LEARNING

KARUN ADUSUMILLI, FRIEDRICH GEIECKE & CLAUDIO SCHILTER

Abstract. Devising guidance on how to assign individuals to treatment is an important goal of empirical research. In practice individuals often arrive sequentially, and the planner faces various constraints such as limited budget/capacity, or borrowing constraints, or the need to place people in a queue. For instance, a governmental body may receive a budget outlay at the beginning of an year, and it may need to decide how best to allocate resources within the year to individuals who arrive sequentially. In this and other examples involving inter-temporal trade-offs, previous work on devising optimal policy rules in a static context is either not applicable, or is sub-optimal. Here we show how one can use offline observational data to estimate an optimal policy rule that maximizes ex-ante expected welfare in this dynamic context. We allow the class of policy rules to be restricted for computational, legal or incentive compatibility reasons. The problem is equivalent to one of optimal control under a constrained policy class, and we exploit recent developments in Reinforcement Learning (RL) to propose an algorithm to solve this. The algorithm is easily implementable and computationally efficient, with speedups achieved through multiple RL agents learning in parallel processes. We also characterize the statistical regret from using our estimated policy rule. To do this, we show that a Partial Differential Equation (PDE) characterizes the evolution of the value function under each policy. The data enables us to obtain a sample version of the PDE that provides estimates of these value functions. The estimated policy rule is the one with the maximal estimated value function. Using the theory of viscosity solutions to PDEs we show that the policy regret decays at a $n^{-1/2}$ rate in most examples; this is the same rate as that obtained in the static case.

This version: December 4, 2019

We would like to thank Tim Armstrong, Debopam Bhattacharya, Xiahong Chen, Denis Chetverikov, Frank Diebold, Wouter Den Haan, John Ham, Toru Kitagawa, Frank Schorfheide, Aleksey Tetenov, Alwyn Young and seminar participants at Brown, University of Pennsylvania, Vanderbilt, Yale, Greater NY area Econometrics conference, Bristol Econometrics study group, and the CEMMAP/UCL workshop on Personalized Treatment for helpful comments.

1. Introduction

Consider a situation wherein a stream of individuals arrive sequentially - e.g, when they get unemployed - to a social planner. Once each individual arrives, our planner needs to decide on an action, i.e a treatment assignment - e.g, whether or not to offer free job training - for the individual, taking into account the individual's characteristics and various institutional constraints such as limited budget/capacity, waiting times and/or borrowing constraints. The decision on the treatment is to be taken instantaneously. It is taken without knowledge of the characteristics of future individuals, though the planner can, and should, form expectations over these future characteristics. Once an action is taken, the individual is assigned a specific treatment, leading to a reward, i.e a change in the utility for that individual. The planner does not observe these rewards directly since they may be only realized much later, but she can estimate them using data from some past observational studies. At the same time, the action of the planner generates an observed change to the institutional variable, such as a reduced budget or increased wait times. The planner takes note of these changes, and waits for the next individual to arrive. This process may repeat indefinitely, or end when some terminal constraints are hit, e.g, when budget or capacity is depleted. In this paper, we propose a Reinforcement Learning algorithm to obtain the welfare maximizing treatment allocation rule for this dynamic setting.

We contend that dynamical constraints are common across governmental and non-governmental settings. The following examples serve to illustrate the generality of our approach:

Example 1.1. (Finite budget) Suppose that a social planner has received an one-off outlay of funds to be allocated to provide treatment to individuals, for example a NGO that has received a single large donation. The planner faces a trade-off in terms of using some of the funds to treat an individual at the moment, or holding off until a more deserving individual arrives in the future. The utility of future individuals is discounted. The planner would like to determine the optimal policy rule for treating individuals. Since the budget declines to 0, the optimal policy rule will be a function of the individual covariates and current budget (and possibly time if we allow for the arrival rates of individuals to vary with time).

Example 1.2. (Infinite horizon and optimal control) Suppose now that the planner receives a steady flow of revenue and individuals arrive at a constant rate, drawn from some underlying distribution that is time-invariant. Ideally, the planner would like to determine a rule for treatment based on individual characteristics so that expected costs equal revenue and the budget stays constant, preferably at a level that is just above 0. Somewhat surprisingly, even in this simple context, a 'static' policy - i.e one which does not change with current budget level - is unsatisfactory. Indeed, under such a policy, the budget would set off on a random walk since the individuals are iid draws from a distribution, and the expected change to budget is 0 only on average. Consequently, the budget would eventually violate any possible borrowing constraint. On the other hand, if the policy were allowed to change with budget, we could find one that varies in just the right way so as to nudge the budget back onto a constant level. Thus a well chosen policy rule allows the planner to achieve some amount of optimal control over the budget process. In this paper we show how one can solve for such a policy rule. In fact we are able to do so under settings more realistic than the one described here and that allow for: (1) the revenue to follow an exogenous process that varies with time, (2) arrival rates of the individuals to vary with time (e.g, due to seasonality in unemployment), (3) the distribution of individuals to change with time (e.g, due to different seasonal trends in unemployment among different groups), and (4) uncertainty in forecasts of arrival rates (e.g, uncertainty in unemployment forecasts).

Example 1.3. (Finite horizon) As a third possibility, suppose that the planner receives an operating budget for each period, e.g a year. Any unused funds will be sent back at the end of the year. Stylized as it is, this setup could serve as a good approximation for how some governmental programs are run in real life, with a budget outlay that is determined by the legislature at the beginning of each financial year. As in the previous example, a static policy is unsatisfactory since it would now lead to the budget process following a random walk with drift.[1] On the other hand, a policy that changes with budget or time allows for the possibility to re-optimize when the budget falls lower or higher than expected, and will thus increase overall welfare. We show how to solve for such a policy. We do this even while allowing for the distribution of individuals to change with time, and while also accounting for uncertainty in the forecasts for the arrival rates. Both situations are again ones in which a static policy would be sub-optimal.

Example 1.4. (Queues) In some situations the planner is constrained not by budget or capacity, but by the amount of time individuals have to wait before getting treatment. This is because the planner needs to expend time to treat an individual, which is much longer than the average time between the arrivals of two individuals. For instance, the treatment could be a medical procedure that takes time, or an unemployment service that requires the individual to meet with a case-worker to help with job applications. In such cases, individuals selected for treatment would be placed in a queue. But waiting is usually costly, and the impact of treatment a decreasing function of the waiting times. Therefore the planner may decide to turn people away from treatment if the length of the queue is too long. As long as the cost of waiting is known or could be estimated using the data, we can use the methods in this paper to determine the optimal rule for whether or not to place an individual in a queue.[2] Such a rule will be a function of the individual characteristics and current length of the queue.

For a related example, suppose there are now two queues, and individuals may be placed in either one depending on their characteristics.[3] The planner could reserve the shorter queue for individuals deemed to be more at risk. She would therefore like a rule to determine which queue to place an individual in, as a function of individual characteristics and the length of both queues. One could again solve for this using our techniques as long as there is some information on the effect of waiting times on different individuals.

[1]So the planner may run out of budget too soon, or is left with too large a budget surplus at the end of the year.

[2]For instance, in many administrative datasets, it is possible to find the duration of the unemployment spell immediately preceding the enrollment into a labor market program, see the analyses of Crepon *et al* (2009), Lechner & Wunsch (2013) and Vikstrom (2017). This duration can be used as a proxy for waiting time.

[3]Something akin to this happens in hospital emergency rooms, though the exact mechanism - whether to use more than two queues etc - is different, see Woodworth and Holmes (2018).

Example 1.5. (Capacity constraints) For our final example, consider capacity constraints. The treatment program might require caseworkers to do home visits, and there are only a fixed number of them who are employed.[4] The planner is thus forced to turn away individuals when the capacity is full.[5] However people finish treatment at some (known or estimable) rate which frees up capacity. The planner would then like to find a treatment rule that allocates individuals to treatment as a function of current capacity and individual covariates.[6]

In all these examples, we show how one can leverage observational data to estimate the optimal policy function that maximizes ex-ante expected welfare. We do this under both full and partial compliance to the policy. Furthermore, we propose algorithms to solve for the optimum within a pre-specified policy class. As explained by Kitagawa and Tetenov (2018), one may wish to restrict the policy class for computational or legal reasons. Another reason is incentive compatibility, e.g, the planner may want the policy to change slowly with time to prevent individuals from manipulating arrival times. The key assumption that we do impose is that the environment, i.e the arrival rates and distribution of individuals, is not affected by the policy. This is a reasonable assumption in settings like unemployment, arrivals to emergency rooms, childbirth (e.g, for provision of daycare) etc., where either the time of arrival is not in complete control of the individual, or it is determined by factors exogenous to the provision of treatment. Alternatively, the planner can employ techniques such as queues that discourage individuals from delaying arrival times. Finally, even where this assumption is suspect, most of our results will continue to apply if we have a model of response to the policy (see Section 6.5).

If the dynamic aspect can be ignored, there exist a number of methods to estimate an optimal policy function that maximizes social welfare, starting from the seminal contribution of Manski (2004), and further extended by Hirano and Porter (2009), Stoye (2009, 2012), Chamberlain (2011), Bhattacharya and Dupas (2012) and Tetenov (2012), among others. More recently, Kitagawa and Tetenov (2018), and Athey and Wager (2018) proposed using Empirical Welfare Maximization (EWM) in this context. While these papers address the question of optimal treatment allocation under co-variate heterogeneity, the resulting treatment rule is static in that it is determined ex-ante, before observing the data on which it will be applied. It does not change with time, nor with current values of institutional constraints. In fact, in some of our examples - Examples 1.1, 1.4, 1.5 - EWM is not even applicable. This is so even if we restricted ourselves to using a static policy. For instance, with budget constraints, the EWM rule requires one to specify the fraction of population that can be treated, but in our dynamic environment the number of individuals the planner faces is endogenous to the policy.

There also exist a number of methods for estimating the optimal treatment assignment policy using 'online' data. This is known as the contextual bandit problem, and there is a large literature on this, see e.g, Dudik *et al* (2011), Agarwal *et al* (2014), Russo and van Roy (2016)

[4]Some examples of programs that require home visits include child FIRST, and the Nurse-Family partnership.

[5]We could consider other alternatives to turning people away, e.g the planner may place individuals in queues. Or, she could hire more caseworkers on a temporary basis, but this comes with some cost.

[6]However, our methods only allow for finite dimensional states. Therefore we are not currently able to accommodate situations wherein the time at which people leave the treatment depends on when they first arrived, since the (infinite dimensional) history of arrivals now becomes a relevant state variable.

and Dimakopoulou *et al* (2017). However, bandit algorithms do not have a forward looking nature; the eventual policy function that is learnt is still static in that it does not take into account the effect of current actions on future states or rewards. By contrast, our primary goal in this paper is to use 'offline', i.e historical data, to estimate a policy rule that is optimal under such inter-temporal trade-offs. But as a by-product, our algorithms can also be applied in a completely offline manner in infinite-horizon Markov Decision Process settings, such as Example 1.1, where the usual bandit algorithms do not apply.[7] In these settings, we can guarantee that our algorithm will eventually learn the optimal policy function, but we do not claim it is welfare-optimal in the interim; see Section 6.4 for more details.

Another close set of results to our work is from the literature on Dynamic Treatment Regimes (DTRs). We refer to Laber *et al* (2014) and Chakraborty and Murphy (2014) for an overview. DTRs consist of a sequence of individualized treatment decisions for health related interventions. These are typically estimated from sequential randomized trials (Murphy, 2005; Lei *et al.*, 2012), where participants move through different stages of treatment, which is randomized in each stage. By contrast, we only make use of a single set of observational data, and this data itself does not come in a dynamic form. Each individual in our setup is only exposed to treatment once. The dynamics are faced not by the individual, but by the social planner. Additionally, in DTRs the number of stages or decision points is quite small, typically between 1 and 5. By contrast, the number of decision points, i.e the rate of arrivals, in our setting is very high, and we will find it more convenient to formulate the model as a differential equation.

In this paper we propose techniques for estimating an optimal policy function that maps the current state variables of observed characteristics and institutional constraints to probabilities over the set of actions. We treat the class of policy functions as given. Then for any policy from that class, we can write down a Partial Differential Equation (PDE) that characterizes the expected value function under that policy, where the expectation is taken over the distribution of the individual covariates. Using the data, we can similarly write down a sample version of the PDE that provides estimates of these value functions. The estimated policy rule is the one that maximizes the estimated value function at the start of the program. By comparing the PDEs, we can uniformly bound the difference in their corresponding solutions, i.e the value functions. This enables us to bound the welfare regret from using the estimated policy rule relative to the optimal policy in the candidate class. We find that the regret is of the (probabilistic) order $n^{-1/2}$ in many cases (Examples 1.1-1.3 & 1.5); this is also the minimax rate for the regret in the static case (see, Kitagawa & Tetenov, 2018). An important requirement for obtaining the $n^{-1/2}$ rate is to employ doubly robust estimates for calculating the rewards (see, Athey and Wager, 2018). As in both these papers, the rate further depends on the complexity of the policy function class being considered, as indexed by its VC dimension.

In the static setting there is a close connection between optimal treatment rules and classification that can be exploited for proving theoretical results and for proposing practical algorithms. In our dynamic setting, the relevant connection is to optimal control. This requires new theoretical methods since there is heavy dependence on the state variables between current and future

[7]See Sutton and Barto (2018, Chapter 3) on the difference between Markov Decision and bandit problems.

periods (e.g, as in Example 1.2, the budget could follow a random walk). Our theory is thus based on exploiting the properties of the PDEs for the expected value functions. Due to the nonlinear nature of the PDEs in our setting, our analysis will be based on the concept of viscosity solutions that allows for non-differentiable solutions, see Crandall, Ishii and Lions (1992) for a survey. This concept has been recently used to analyze heterogeneous agent macroeconomic models in continuous time (Achdou et al, 2018).

In terms of computation, we approximate the PDEs with suitable dynamic programming problems by discretizing the number of arrivals. We then propose a modified Reinforcement Learning algorithm that can be applied on the latter and that achieves the best value in a pre-specified class of policy rules. Previous work in this literature in economics has used Generalized Policy Iteration (e.g, Benitez-Silva *et al*, 2000). While this method works well with discrete states, there are three major drawbacks: First, and most importantly, it does not allow for restricting the solution to a pre-specified class of policy rules. Second, the algorithm becomes cumbersome even with a few continuous states, and a few thousand decision points.[8] Third, it cannot be directly applied to our setup without incorporating a regularization parameter to avoid over-fitting the value function (and it is not obvious how such a regularization may be employed). This is because standard reward estimates (inverse propensity weighting, doubly robust etc.) are direct functions of the outcome variables from the observational data. Hence the usual policy iteration algorithm would overfit the estimate of the value function to this data. In this paper, we propose a modified Reinforcement Learning (RL) algorithm that solves all these issues.[9] We adapt the Actor-Critic algorithm (e.g Sutton *et al*, 2000; Bhatnagar *et al*, 2009) that has been applied recently to great effect in applications as diverse as playing Atari games (Mnih *et al*, 2015), image classification (Mnih *et al*, 2014) and machine translation (Bahdanau *et al*, 2016). Our algorithm avoids the over-fitting issue by working with the expected value function that integrates over the rewards at each step. The integration is implicit since we use stochastic gradient descent, so the computational complexity is not affected.

Our Reinforcement Learning algorithm appears to be a novel approach to the solution of Hamilton-Jacobi-Bellman type PDEs. In addition to possessing strong convergence properties, it is also parallelizable, which translates to very substantial computational gains. We also outline the computational and numerical properties of our algorithm. On the computational side, we prove that it converges to a well defined optimum. This is based on the convergence of stochastic gradient descent, and we are able to directly employ theorems from the RL literature to this effect. On the numerical approximation side, we use results from the theory of viscosity solutions to provide conditions on the level of discretization so that the numerical error from this is negligible compared to the statistical error in the regret bounds.

We illustrate the feasibility of our algorithm using data from the Job Training Partnership Act (hereafter JTPA). We incorporate dynamic considerations into this setting in the sense that

[8]Continuous states may be handled through discretization or parametric policy iteration. The former is typically slower and suffers from a strong curse of dimensionality (see Benitez-Silva *et al*, 2000, Section 2.5); while the latter requires numerical integration which is also very demanding with more than a few states. Also, there is no proof of convergence for parametric policy iteration, and it is known that it fails to converge in some examples.

[9]We refer to Sutton and Barto (2018) for a detailed comparison of recent RL algorithms with policy iteration.

the planner has to choose whether to send individuals for training as they arrive sequentially. The planner faces a budget constraint, and the population distribution of arrivals is also allowed to change with time. We consider policy rules composed of 5 continuous state variables (3 individual covariates along with time and budget), to which we add some interaction terms. We then apply our Actor-Critic algorithm to estimate the optimal policy rule.

2. An illustrative example: Dynamic treatment allocation with a finite budget constraint

To illustrate our setup and methods consider the following simplified version of Example 1: A social planner wants to provide training to unemployed people. The planner starts with a fixed budget that she can use to fund the training. Individuals arrive sequentially when they get unemployed, and the planner is required to provide an instantaneous decision on whether to provide training to the current individual, or to to hold off for a more eligible applicant at the risk of losing some utility due to discounting. The decision may be based on the current budget and the characteristics of the individual. To help with the decision, the planner can draw on information from a historical Randomized Control Trial (RCT) on the effect of training, along with data on unemployment dynamics. The program ends when the budget is depleted. We assume in this section that the waiting time between arrivals is drawn from an exponential distribution with a constant parameter (i.e the unemployment rates are assumed to not change with time), and also that the cost of training is the same for all individuals. This allows us to characterize the problem in terms of Ordinary Differential Equations (ODEs), which greatly simplifies the analysis. We consider more general setups and other examples, leading to Partial Differential Equations (PDEs), in the next section.

Formally, let x denote the vector of characteristics of an individual, based on which the planner makes a decision on whether to provide training ($a = 1$) or not ($a = 0$). The current budget is denoted by z. Once an action, a, has been chosen, the planner affects an increase in social welfare by the quantity $Y(a)$ that is equivalent to the potential outcome of the individual under action a. We shall assume for this section that $Y(a)$ is affected by the covariates x but not the budget. Define $r(x, a) = E[Y(a)|x]$ as the expected (instantaneous) reward for the social planner when the planner chooses action a for an individual with characteristics x. Since we only consider additive welfare criteria in this paper, we may normalize $r(x, 0) = 0$, and set $r(x, 1) = E[Y(1) - Y(0)|x]$. Note that we can accommodate various welfare criteria, as long as they are utilitarian, by redefining the potential outcomes.

If the planner takes action $a = 1$, her budget is depleted by c, otherwise it stays the same. The next individual arrives after a waiting time Δt drawn from an exponential distribution $\text{Exp}(N)$. Note that N is the expected number of individuals in a time interval of length 1 (one could alternatively use this as the definition of N itself). We shall use N to rescale the budget so that $c = 1/N$. With this, we reinterpret the budget as the fraction of people in a unit time period that can be treated. Each time a new individual arrives, the covariates for the individual are assumed to be drawn from a distribution F that is fixed but unknown. The utility from treating successive individuals is discounted exponentially by $e^{-\beta \Delta t}$. Note that the expected discount factor is given by $E[e^{-\beta \Delta t}] = 1 - \frac{\tilde{\beta}}{N}$ where $\tilde{\beta} = \beta + O(N^{-1})$. For simplicity, we shall let $\tilde{\beta} = \beta$.

The planner chooses a policy function $\pi(a|x,z)$ that maps the current state variables x, z to a probabilistic choice over the set of actions:

$$\pi(.|x,z) : (x,z) \longrightarrow [0,1].$$

The aim of the social planner is to determine a policy rule that maximizes expected welfare after discounting. Let $v_\pi(x,z)$ denote the value function for policy π, defined as the per-person expected welfare from implementing policy $\pi(\cdot|x,z)$ starting from the state (x,z). In other words,

$$v_\pi(x,z) = E\left[\frac{1}{N} \sum_{i=1}^{\infty} e^{-\beta T_i} r(x_i,1)\pi(1|x_i,z_i)\mathbb{I}(z_i > 0) \middle| x,z \right],$$

where the expectation is joint over the times of arrival T_i of each individual, covariates $x \sim F$ and z_i which evolves according to the distribution of x and the randomization of the policy $\pi(.)$. We can also represent $v_\pi(z,t)$ in a recursive form as the fixed point to the equations[10]

$$\begin{aligned} v_\pi(x,z) &= \frac{r(x,1)}{N}\pi(1|x,z) + \left(1 - \frac{\beta}{N}\right) E_{x' \sim F}\left[v_\pi\left(x', z - \frac{1}{N}\right)\pi(1|x,z) + v_\pi(x',z)\pi(0|x,z)\right] \\ &\quad \text{for } z > 1/N \\ v_\pi(x,0) &= 0. \end{aligned}$$

To obtain a more insightful expression, we can integrate out x. This motivates the integrated value function:

$$h_\pi(z) := E_{x\sim F}[v_\pi(x,z)].$$

Define $\bar{\pi}(a|z) = E_{x\sim F}[\pi(a|x,z)]$ and $\bar{r}_\pi(z) = E_{x\sim F}\left[r(x,1)\pi(1|x,z)\right]$. We can then characterize $h_\pi(.)$ as the solution to the recursive equations

$$\begin{aligned} h_\pi(z) &= \frac{\bar{r}_\pi(z)}{N} + \left(1 - \frac{\beta}{N}\right)\left\{ h_\pi\left(z - \frac{1}{N}\right)\bar{\pi}(1|z) + h_\pi(z)\bar{\pi}(0|z)\right\} \text{ for } z > 1/N, \\ h_\pi(0) &= 0. \end{aligned} \tag{2.1}$$

In practice the value of N is very large, i.e the rate of arrival of people is very fast, so that budget is almost continuous. In such cases it is more convenient to work with the limiting version of (2.1) as $N \to \infty$. To this end let us subtract $\left(1 - \frac{\beta}{N}\right) h_\pi(z)$ from both sides of equation (2.1), multiply both sides by N and take the limit as $N \to \infty$. We then end up with the following Ordinary Differential Equation (ODE) for the evolution of $h_\pi(.)$:

$$\beta h_\pi(z) = \bar{r}_\pi(z) - \bar{\pi}(1|z)\partial_z h_\pi(z), \qquad h_\pi(0) = 0, \tag{2.2}$$

where ∂_z denotes the differential operator with respect to z.[11] ODE (2.2) is similar to the well known Hamilton-Jacobi-Bellman (HJB) equation. However, an important difference is that (2.2) determines the evolution of $h_\pi(.)$ under a specified policy, while the HJB equation determines the evolution of the value function under the optimal policy.

[10]We assume for simplicity that z is always in multiples of $1/N$.

[11]Sufficient conditions for a unique solution to (2.2) are provided in Appendix C.3.

Note that ODE (2.2) also provides a very good approximation to the value function if, e.g, the planner groups all individuals arriving in a day and employs the same policy function (i.e keeping z, t fixed) for all of them. We need not restrict the planner to considering individuals only one at a time. Changes to z, t are negligible if the numbers in the groups are small compared to the number of people being considered overall. We also note that there is an alternative way in which we could have 'derived' the same ODE: This is if we discretized time into periods, and assumed that number of people who arrive in each period is a Poisson random variable with parameter $\lambda(t)\Delta l$, where Δl denotes the time step (days, minutes etc) between two successive periods. We would then obtain ODE (2.2) in the limit as $\Delta l \to 0$.

The social planner's decision problem is to choose the optimal policy π^* that maximizes the ex-ante expected welfare $h_\pi(z_0)$, over a pre-specified class of policies Π, where z_0 denotes the initial value of the budget:

$$\pi^* = arg \max_{\pi \in \Pi} h_\pi(z_0).$$

How should the planner choose Π? Consider the first best policy function:

$$\pi^*_{FB}(1|x, z) = \mathbb{I}\left\{r(x, 1) - \partial_z h_{\pi^*_{FB}}(z) > 0\right\}.$$

In general settings it is not clear what, if any, regularity properties $\pi^*_{FB}(.)$ possesses. In the next section, we will find that once we have time as a state variable, ODE (2.2) becomes a PDE, and the value function is generically non-differentiable, which makes characterizing $\pi^*_{FB}(\cdot)$ difficult. Consequently, it is not clear that consistent estimation of $\pi^*_{FB}(1|x, z)$ is possible with sample data. Furthermore, $\pi^*_{FB}(1|x, z)$ would be generally be discontinuous and highly non-linear in (x, z), and the social planner may prefer policies that are simpler for legal, ethical or incentive compatibility reasons. For instance, if the policy function is discontinuous in z, individuals may decide to arrive at slightly different times where the budget is different. Ultimately, the choice of Π depends on computational and policy considerations of the planner. For our theoretical results we take this as given and consider a class Π of policies indexed by some possibly infinite dimensional parameter $\theta \in \Theta$. For instance, Θ could represent a collection of sets or functions.

For computation, however, we require $\pi_\theta(.)$ to be differentiable in θ. This still allows for rich spaces of policy functions. A rather convenient one is the class of soft-max functions. Let $f(x, z)$ denote a vector of functions of dimension k. The soft-max function takes the form

$$\pi_\theta^{(\sigma)}(1|x, z) = \frac{\exp(\theta^\intercal f(x, z)/\sigma)}{1 + \exp(\theta^\intercal f(x, z)/\sigma)}. \tag{2.3}$$

Here, θ is normalized by setting one of the coefficients, e.g the intercept, to 1. The term σ is a 'temperature' parameter that is either determined beforehand, or estimated along with θ, in which case it could be subsumed into the latter. For a fixed σ, we define the soft-max policy class as $\Pi_\sigma := \{\pi_\theta^{(\sigma)}(\cdot|s) : \theta \in \Theta\}$. As $\sigma \to 0$, this becomes equivalent to the class of generalized linear eligibility scores proposed by Kitagawa and Tetenov (2018), which are of the form $\mathbb{I}\{\theta^\intercal f(x, z) > 0\}$. More generally, the class $\{\pi_\theta^{(\sigma)}(1|x, z) : \theta \in \mathbb{R}^k, \sigma \in \mathbb{R}^+\}$ can approximate any deterministic policy, including $\pi^*_{FB}(.)$, arbitrarily well, given a large enough dimension k. For even more expressive policies, this can be generalized, e.g to multi-layer neural networks.

Note that for computation we cannot directly work with deterministic rules as they are not differentiable in θ. In practice, however, we can employ the soft max class and let $\sigma \to 0$ in the course of computation to obtain a deterministic rule, see Section 6.3. Alternatively, we can let σ be estimated, and the algorithm will converge to a deterministic policy if the latter is optimal.

In what follows, we specify the policy class as $\Pi \equiv \{\pi_\theta(.) : \theta \in \Theta\}$ and denote $h_\theta \equiv h_{\pi_\theta}$ along with $\bar{r}_\theta \equiv \bar{r}_{\pi_\theta}$. The social planner's problem is then

$$\theta^* = arg \max_{\theta \in \Theta} h_\theta(z_0). \tag{2.4}$$

Clearly (2.4) is not feasible as one does not know $r(x, 1)$, nor the distribution F to calculate $h_\theta(z)$. However the planner does have access to an RCT, which we assume to consist of an iid draw of size n from the distribution F. The empirical distribution F_n of these observations is thus a good proxy for F. Let W denote the treatment assignment in the RCT data. We also let $\mu(x, w) = E[Y(w)|X = x, W = w]$ denote the conditional expectations for $w = 0, 1$, and $p(x)$, the propensity score. We recommend a doubly robust method to estimate $r(x, 1)$, e.g,

$$\hat{r}(x, 1) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0) + (2W - 1)\frac{Y - \hat{\mu}(x, W)}{W\hat{p}(x) + (1 - W)(1 - \hat{p}(x))},$$

where $\hat{\mu}(x, w)$ and $\hat{p}(x)$ are non-parametric estimates of $\mu(x, w)$ and $p(x)$ respectively, and Y is the observed outcome variable.

Define $\hat{\pi}_\theta(a|z) = E_{x\sim F_n}[\pi_\theta(a|x, z)]$ and $\hat{r}_\theta(z) = E_{x\sim F_n}[r(x, 1)\pi_\theta(1|x, z)]$. Based on the knowledge of $\hat{r}(.)$ and F_n, we can calculate a sample estimate of the integrated value function in the discrete case as the solution to

$$\hat{h}_\theta(z) = \frac{\hat{r}_\theta(z)}{N} + \left(1 - \frac{\beta}{N}\right)\left\{\hat{h}_\theta\left(z - \frac{1}{N}\right)\hat{\pi}_\theta(1|z) + \hat{h}_\theta(z)\hat{\pi}_\theta(0|z)\right\} \text{ for } z > 1/N, \tag{2.5}$$
$$\hat{h}_\theta(0) = 0.$$

Alternatively, in the limit as $N \to \infty$, we have the following ODE:

$$\beta\hat{h}_\theta(z) = \hat{r}_\theta(z) - \hat{\pi}_\theta(1|z)\partial_z\hat{h}_\theta(z), \qquad \hat{h}_\theta(0) = 0. \tag{2.6}$$

Using $\hat{h}_\theta(.)$ we can solve a sample version of the social planner's problem:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \hat{h}_\theta(z_0).$$

Given θ, one could solve for $\hat{h}_\theta$ by backward induction starting from $z = 1/N$ using (2.5). In this simple example this is feasible as long as N is not too large, but note that one would need to calculate the quantities $E_{x\sim F_n}[\pi_\theta(a|x, z)]$ and $E_{x\sim F_n}[r(x, 1)\pi_\theta(1|x, z)]$ - which are averages over n observations - for all possible values of z. And even where solving for $\hat{h}_\theta(z_0)$ is feasible, we yet have to maximize this over $\theta \in \Theta$. Such a strategy is computationally too demanding. Therefore in this paper we advocate a Reinforcement Learning algorithm that directly ascends along the gradient of $\hat{h}_\theta(z_0)$ *and* simultaneously calculates $\hat{h}_\theta(z_0)$ in the same series of steps. Furthermore, in making use of stochastic gradient descent, the algorithm only samples the quantities $E_{x\sim F_n}[\pi_\theta(a|x, z)]$ and $E_{x\sim F_n}[r(x, 1)\pi_\theta(1|x, z)]$, instead of taking averages. All this makes the algorithm very efficient. We describe our algorithm in Section 4.

In the remainder of this section, we briefly outline the theory behind our approach. The derivations here are informal, but provide intuition for our formal results in Section 5.

2.1. **Regret bounds.** We would like to know how $\hat{\theta}$ compares to θ^* in terms of the welfare regret $h_{\theta^*}(z_0) - h_{\hat{\theta}}(z_0)$. The bound for this depends on the sample size n and the complexity of the space $\Pi = \{\pi_\theta : \theta \in \Theta\}$. One way to determine the complexity of Π is by its Vapnik-Cervonenkis (VC) dimension. In particular, denote by v the VC-subgraph index of the collections of functions

$$\mathcal{I} = \{\pi_\theta(1|\cdot, z) : z \in [0, z_0], \theta \in \Theta\}$$

indexed by z and θ. We shall assume that v is finite. Kitagawa and Tetenov (2018) were the first to characterize the regret in the static setting in terms of the VC dimension of Π. Relative to this, our definition of the complexity differs in two respects. First, our policy functions are probabilistic. Second, for the purposes of calculating the VC dimension, we treat z as an index to the functions $\pi_\theta(1|\cdot, z)$, similarly to θ. In other words $\pi_\theta(1|\cdot, z_1)$ and $\pi_\theta(1|\cdot, z_2)$ with the same θ are treated as different functions. This is intuitive since how rapidly the policy rules change with budget is also a measure of their complexity. Note that the VC index of $\mathcal{I}$ is not $\dim(\theta)$ when θ is Euclidean, but is in fact smaller. To illustrate, suppose that x is univariate and

$$\mathcal{I} \equiv \{\text{Logit}(g_1(z) + g_2(z)x) : g_1, g_2 \text{ are arbitrary functions}\}.$$

In this case the VC-subgraph index of $\mathcal{I}$ is at most 2.[12]

We now show how one can derive probabilistic bounds for the regret $h_{\theta^*}(z_0) - h_{\hat{\theta}}(z_0)$. First, under the assumption of finite VC dimension and other regularity conditions, Athey and Wager (2018) show that for doubly robust estimates of the rewards,

$$E_{x\sim F}\left[\sup_{\theta\in\Theta, z\in[0,z_0]} |\bar{r}_\theta(z) - \hat{r}_\theta(z)|\right] \le C_0\sqrt{\frac{v}{n}}, \tag{2.7}$$

$$E_{x\sim F}\left[\sup_{\theta\in\Theta, z\in[0,z_0]} |\bar{\pi}_\theta(1|z) - \hat{\pi}_\theta(1|z)|\right] \le C_0\sqrt{\frac{v}{n}}$$

for some universal constant $C_0 < \infty$. Denote $\hat{\delta}_\theta(z) = h_\theta(z) - \hat{h}_\theta(z)$. Now under some regularity conditions (made precise in Section 5), it can be shown that $\sup_{\theta\in\Theta, z\in[0,z_0]} |h_\theta(z)| < \infty$. Then from (2.2) and (2.6), we have

$$\partial_z\hat{\delta}_\theta(z) = \frac{-1}{\bar{\pi}_\theta(1|z)}\beta\hat{\delta}_\theta(z) + \frac{\bar{r}_\theta(z)}{\bar{\pi}_\theta(1|z)} - \frac{\hat{r}_\theta(z)}{\hat{\pi}_\theta(1|z)} + \left(\frac{1}{\hat{\pi}_\theta(1|z)} - \frac{1}{\bar{\pi}_\theta(1|z)}\right)\beta\hat{h}_\theta(z); \quad \hat{\delta}_\theta(0) = 0$$

or

$$\partial_z\hat{\delta}_\theta(z) = \frac{-1}{\bar{\pi}_\theta(z)}\beta\hat{\delta}_\theta(z) + K_\theta(z); \quad \hat{\delta}_\theta(0) = 0, \tag{2.8}$$

where

$$E_{x\sim F}\left[\sup_{\theta\in\Theta, z\in[0,z_0]} |K_\theta(z)|\right] \le M\sqrt{v/n}$$

[12]To see this, note that the VC-subgraph index of the class of functions $\mathcal{F} = \{f : f(x) = a + xb \text{ over } a, b \in \mathbb{R}\}$ is 2 since $\mathcal{F}$ lies in the (two dimensional) vector space of the functions $1, x$. The VC-subgraph index of $\mathcal{I}$ is the same or lower than that of $\mathcal{F}$ (since the logit transformation is monotone), hence $v \le 2$ in this example.

for some $M < \infty$ by (2.7) and the uniform boundedness of $h_\theta(z)$, assuming that $\bar{\pi}_\theta(z)$ is uniformly bounded away from 0. Rewriting (2.8) in integral form and taking the modulus on both sides, we obtain

$$\left|\hat{\delta}_\theta(z)\right| \le zM\sqrt{\frac{v}{n}} + \int_0^z \frac{1}{\bar{\pi}_\theta(\omega)}\beta\left|\hat{\delta}_\theta(\omega)\right| d\omega,$$

based on which we can conclude via Grönwall's inequality that

$$\left|\hat{\delta}_\theta(z)\right| \le M_1\sqrt{\frac{v}{n}}$$

uniformly over all $\theta \in \Theta, z \in [0, z_0]$, for some $M_1 < \infty$ - here, all the inequalities should be interpreted as holding with probability approaching one under F. The above discussion implies

$$h_{\theta^*}(z_0) - h_{\hat{\theta}}(z_0) \le 2 \sup_{\theta\in\Theta, z\in[0,z_0]} \left|\hat{\delta}_\theta(z)\right| \le 2M_1\sqrt{\frac{v}{n}}$$

with probability approaching one under F. Hence the regret declines with $\sqrt{v/n}$, which is the same rate that Kitagawa and Tetenov (2018) derived for the static case.

2.2. **Discretization and numerical error.** As we mentioned earlier, we do not recommend using the ODE version of the problem to solve for $\hat{\theta}$. Instead, it is usually much quicker to solve a discrete analogue of the problem as in (2.5). Now in practice N maybe unknown or too large, but in either case we can simply employ any suitably large normalizing factor b_n, and solve the recurrence relation

$$\tilde{h}_\theta(z) = \frac{\bar{r}_\theta(z)}{b_n} + \left(1 - \frac{\beta}{b_n}\right)\left\{\tilde{h}_\theta\left(z - \frac{1}{b_n}\right)\bar{\pi}_\theta(1|z) + \tilde{h}_\theta(z)\bar{\pi}_\theta(0|z)\right\} \tag{2.9}$$

for $\tilde{h}_\theta(.)$ together with the initial condition $\tilde{h}_\theta(0) = 0$. We are now faced with the issue of choosing b_n so that $\tilde{h}_\theta(.)$ is sufficiently close to $\hat{h}_\theta(.)$ obtained from (2.6).

To answer this, we first note that $\hat{h}_\theta$ and $\partial_z\hat{h}_\theta$ are both Lipschitz continuous uniformly in θ under some regularity conditions (c.f Section 5). Lipschitz continuity of $\partial_z\hat{h}_\theta$ implies

$$\hat{h}_\theta(z) = \frac{\bar{r}_\theta(z)}{b_n} + \left(1 - \frac{\beta}{b_n}\right)\left\{\hat{h}_\theta\left(z - \frac{1}{b_n}\right)\bar{\pi}_\theta(1|z) + \hat{h}_\theta(z)\bar{\pi}_\theta(0|z)\right\} + \frac{B_\theta(z)}{b_n^2},$$

where $|B_\theta(z)| \le B < \infty$ uniformly over θ and z. Then defining $\tilde{\delta}_\theta(z) = \hat{h}_\theta(z) - \tilde{h}_\theta(z)$, and subtracting (2.9) from the previous display equation, we get

$$\tilde{\delta}_\theta(z) = \left(1 - \frac{\beta}{b_n}\right)\left\{\tilde{\delta}_\theta\left(z - \frac{1}{b_n}\right)\bar{\pi}_\theta(1|z) + \tilde{\delta}_\theta(z)\bar{\pi}_\theta(0|z)\right\} + \frac{B_\theta(z)}{b_n^2}.$$

Now let $\mathcal{Z}(n) = \{1/b_n, 2/b_n, \ldots, z_0\}$. From the previous display equation, it follows

$$\sup_{\theta\in\Theta, z\in\mathcal{Z}_n} |\tilde{\delta}_\theta(z)| \le \left(1 - \frac{\beta}{b_n}\right) \sup_{\theta\in\Theta, z\in\mathcal{Z}_n} |\tilde{\delta}_\theta(z)| + \frac{B}{b_n^2},$$

which implies $\sup_{\theta\in\Theta, z\in\mathcal{Z}_n} |\tilde{\delta}_\theta(z)| \le B/b_n$ upon rearrangement. So far $\tilde{h}_\theta(.)$ was only defined for multiples of b_n, but we can extend it to all of $[0, z_0]$ by setting $\tilde{h}_\theta(z) = \tilde{h}_\theta\left(b_n \lfloor z/b_n \rfloor\right)$. Combining the above with the (uniform) Lipschitz continuity of $\hat{h}_\theta(\cdot)$, we obtain

$$\sup_{\theta\in\Theta, z\in[0,z_0]} \left|\tilde{\delta}_\theta(z)\right| = O\left(\frac{1}{b_n}\right).$$

Suppose that θ were estimated using (2.9) as

$$\tilde{\theta} = \arg\max_{\theta \in \Theta} \tilde{h}_\theta(z_0).$$

Then in view of the previous discussion,

$$h_{\theta^*}(z_0) - h_{\tilde{\theta}}(z_0) \leq 2M_1\sqrt{\frac{v}{n}} + 2 \sup_{\theta \in \Theta, z \in [0, z_0]} \left| \tilde{\delta}_\theta(z) \right| = 2M_1\sqrt{\frac{v}{n}} + O\left(\frac{1}{b_n}\right).$$

Hence the numerical error from discretization declines at the rate b_n^{-1}. As long as b_n is chosen to be substantially bigger than $\sqrt{n}$, this approximation error is dwarfed by the statistical error from the regret bound.

3. General setup

In this section, we tackle the question of dynamically optimal treatment allocation in a general setting that nests Examples 1-5 in Section 1 as special cases. The starting point of our approach is a PDE that models the evolution of the social planner's welfare. The different examples from Section 1 will then correspond to various boundary conditions for the PDE. By way of a motivation, we shall start by describing a particular setting, based on a Poisson point process for the arrivals, from which the PDE can be recovered in the limit. We note, however, that this is not only way in which one could motivate the PDE; we discuss other possibilities shortly.

With the above in mind, consider the following setting: The state variables are given by

$$s := (x, z, t),$$

where x denotes the vector of characteristics or covariates of the individual, z is the institutional variable (e.g, the current budget, capacity, or queue length), and t is time. For convenience, we shall take z to be scalar for the rest of this paper. Examples, and extensions to multivariate z can be found in Appendix C.

The arrivals are determined by an inhomogenous Poisson point process with parameter $\lambda(t)N$. Here N is a scale parameter that determines the rate at which individuals arrive, while $\lambda(t)$ itself is normalized via $\lambda(t_0) = 1$. Thus $\lambda(t)$ is the relative frequency of arrivals at time t compared to that at time t_0. As in Section 2, we shall eventually let $N \to \infty$ to end up with a Partial Differential Equation (PDE). We shall also treat $\lambda(t)$ as a forecast and condition on it. For now we focus on a single forecast. But our methods can accommodate multiple forecasts and uncertainty over them. We discuss this in more detail at the end of this section.

As in Section 2, the planner has to choose among actions $a = \{0, 1\}$. The choice of the action is determined by a policy function, $\pi_\theta(a|s)$, indexed by θ:

$$\pi_\theta(.|s) : s \longrightarrow [0, 1].$$

If an action, a, has been chosen, the planner receives a utility of $Y(a)/N$. Observe that, as in Section 2, we have normalized the individual utilities by N. We now allow $Y(a)$ to be affected by all the state variables $s = (x, z, t)$; this extension is needed for Example 1.4 on queues. The rewards are defined as $r(s, 1) = E[Y(1) - Y(0)|s]$, and we set the normalization $r(s, 0) = 0$.

Conditional on the action a and state s, we define an equation, or 'law of motion', to govern the evolution of z as:

$$z' - z = G_a(s)/N,$$

where $G_a(\cdot); a \in \{0,1\}$ is some known function. For example, in the setup of Section 2,

$$G_a(s) = \begin{cases} -1 & \text{if } a = 1 \text{ and } z > 0 \\ 0 & \text{if } a = 0. \end{cases} \tag{3.1}$$

Let us define the quantity $m(t) = \int_{t_0}^{t} \lambda(u)du$, to be interpreted as the expected mass of individuals until time t, when each individual is assigned a weight of $1/N$. We can then reinterpret many of the variables described above as flow quantities relative to m, using the latter as the running or 'base' variable. For instance, $G_a(s)$ can be thought of as flow rate of budget with respect to m. Indeed, suppose that the social planner chooses action a for an infinitesimal mass, δm, of individuals, all with the same covariate x. This corresponds to $N\delta m$ individuals. Then the infinitesimal change to z is given by $\delta z \approx G_a(s)(N\delta m)/N = G_a(s)\delta m$. In a similar vein, we can think of $Y(a)$ and $r(s,a)$ as the flow utilities and flow rewards with respect to m. This interpretation of $G_a(s)$, $Y(a)$ and $r(s,a)$ - as flow quantities - is useful since it is not affected by the normalization or other specifics of the current setting.

Finally, the distribution of the covariates is given by

$$x \sim F,$$

where F is fixed and assumed to not change with t or z.

We left out some extensions for ease of exposition. First, we did not allow the distribution F of the individual covariates to vary with time. In Section 6.2 we relax this using clusters. Second, we have not accommodated the possibility of non-compliance. This is discussed in Section 6.1. We do however maintain the key economic assumption that individuals do not strategically respond to the social planner's policy, e.g, by arriving at different times. Indeed, the waiting times and distribution of covariates were assumed to be independent of all state variables (except for time, we allow this in Section 6.2). We will return to this in Section 6.5.

Define the quantities

$$\bar{r}_\theta(z,t) := E_{x\sim F}[r(s,1)\pi_\theta(1|s)|z,t],$$

and

$$\bar{G}_\theta(z,t) := E_{x\sim F}\left[G_1(s)\pi_\theta(1|s) + G_0(s)\pi_\theta(0|s)|z,t\right].$$

Let $h_\theta(z,t)$ denote the integrated value function. As $N \to \infty$, the evolution of $h_\theta(z,t)$ is determined by the following Partial Differential Equation (PDE):

$$\beta h_\theta(z,t) - \lambda(t)\bar{G}_\theta(z,t)\partial_z h_\theta(z,t) - \partial_t h_\theta(z,t) - \lambda(t)\bar{r}_\theta(z,t) = 0 \text{ on } \mathcal{U}. \tag{3.2}$$

Here $\mathcal{U}$ is the domain of the PDE (more on this below). In Appendix C.1, we show how one can interpret or 'derive' (3.2) in three different ways: (1) as the culmination of a 'no-arbitrage' argument, (2) as the limit of a sequence of discrete dynamic programming problems; and (3) as the characterization of the value function when the arrivals are given by a Poisson point process with parameter $\lambda(t)N$, and $N \to \infty$ (which was the setting so far in this section). In fact, for the last interpretation, we can even set $N = 1$ if the setup is an infinite horizon one and there is no boundary condition on z. The second interpretation is similar to the setup in Section 2. In the next section, we will do the converse, i.e we will approximate the PDE with a discrete dynamic programming problem as a device for computation. The formal justification for this is provided by Theorem 3. In addition, the proof technique for Theorem 3 can be used to formally justify the third interpretation as well, though we omit the details since they are similar.

To complete the dynamic model, we need to specify a boundary condition for (3.2). We consider the different possibilities below:

Dirichlet boundary condition. Under this heading we consider boundary conditions of the form $h_\theta(z,T) = 0 \; \forall z$ (e.g, a finite time constraint), or $h_\theta(\underline{z},t) = 0 \; \forall t$ (e.g, a budget constraint), or both. The quantities $\underline{z}$ and T are some known constants e.g, denoting budget and time constraints. Formally, the set $\mathcal{U}$ is of the form $\mathcal{U} \equiv (\underline{z},\infty) \times [t_0,T)$,[13] and the boundary condition specified as

$$h_\theta(z,t) = 0 \text{ on } \Gamma, \tag{3.3}$$

where $\Gamma \subseteq \partial\mathcal{U}$ is given by

$$\Gamma \equiv \{\{\underline{z}\} \times [t_0,T]\} \cup \{(\underline{z},\infty) \times \{T\}\}. \tag{3.4}$$

Both $\underline{z} = -\infty$ or $T = \infty$ are allowed.

Periodic boundary condition. Consider a setting where the program continues indefinitely. Then t is a relevant state variable only as it relates to some periodic or repeated quantity, e.g seasonality. So, in this setting, $\mathcal{U} \equiv \mathbb{R} \times [t_0,\infty)$, and we impose the periodic boundary condition:

$$h_\theta(z,t) = h_\theta(z,t+T_p) \; \forall (z,t) \in \mathbb{R} \times [t_0,\infty). \tag{3.5}$$

Here, T_p is a known quantity denoting the period length (e.g, a year). Note that the periodic boundary condition can only be valid as long PDE (3.2) is also periodic, i.e the coefficients $\lambda(t), \bar{G}_\theta(z,t), \bar{r}_\theta(z,t)$ are periodic in t with period length T_p. The latter implies that the policy π_θ should also be periodic.

Neumann boundary condition. To motivate this boundary condition, consider the setup of Example 1.3, with a no-borrowing constraint. The social planner is unable provide any treatment when $z = \underline{z} := 0$. Assume that the planner receives a steady flow of funds given by $\sigma(z,t)$, where the flow is defined with respect to time. Then at $z = \underline{z}$, we have $\lambda(t)\bar{G}_\theta(\underline{z},t) = \sigma(\underline{z},t)$ and $\bar{r}_\theta(\underline{z},t) = 0$. Thus (3.2) takes the form

$$\beta h_\theta(z,t) - \sigma(z,t)\partial_z h_\theta(z,t) - \partial_t h_\theta(z,t) = 0, \quad \text{on } \{\underline{z}\} \times [t_0,T). \tag{3.6}$$

[13] We depart slightly here from the usual convention of taking $\mathcal{U}$ to be an open set. In that case $\mathcal{U} \equiv (z_c,\infty) \times (t_0,T)$ but there is no boundary condition at t_0. Since the solution will be continuous, we can always extend it to $t = t_0$, and a short argument will show that (3.2) also holds at t_0 (see e.g, Crandall, Evans and Lions, 1984, Lemma 4.1).

Since (3.6) specifies how the solution behaves at the boundary, we can use it as a boundary condition. Indeed, (3.6) behaves like a reflecting boundary condition since it serves to push the value of z back up when it hits $\underline{z}$.[14]

The chief utility of the boundary condition (3.6) is in allowing the dynamics at the boundary to be different from the interior. Apart from modeling borrowing constraints, this can also be useful in examples with queues or capacity constraints where the end points (e.g when the queue length is 0, or the capacity is full) are treated differently by the social planner. To allow for all this, we consider a setting with a time constraint, and a semi-linear boundary condition on z. Formally, we set $\mathcal{U} \equiv (\underline{z}, \infty) \times [t_0, T)$ and specify the boundary condition to be

$$
\begin{aligned}
(3.7) \qquad \beta h_\theta(z,t) - \bar{\sigma}_\theta(z,t)\partial_z h_\theta(z,t) - \partial_t h_\theta(z,t) - \bar{\eta}_\theta(z,t) &= 0, \quad \text{on } \{\underline{z}\} \times [t_0, T), \\
h_\theta(z,T) &= 0, \quad \text{on } (\underline{z}, \infty) \times \{T\}.
\end{aligned}
$$

Here $\bar{\sigma}_\theta(\underline{z},t)$ and $\bar{\eta}_\theta(z,t)$ are known functions, which are basically the values $\lambda(t)\bar{G}_\theta(s), \lambda(t)\bar{r}_\theta(z,t)$ take on at the boundary $z = \underline{z}$, if the latter were allowed to be discontinuous. The key requirement here is $\bar{\sigma}_\theta(\underline{z},t) > \delta > 0$ for all t. Barles and Lions (1991) term the first part of (3.7) a semi-linear Neumann boundary condition. In general, the Neumann boundary condition may be over-determined and one would have to allow that it may not hold at some points of $\partial\mathcal{U}$. It is therefore important to interpret (3.7) in a viscosity sense, which takes care of these possibilities. We refer the reader to Appendix A for a precise definition.

Periodic Neumann boundary condition. This is an infinite horizon version of the previous case. Suppose that PDE (3.2) is periodic in t with period length T_p. Then, setting $T \to \infty$ in the previous case, we have $\mathcal{U} \equiv (\underline{z}, \infty) \times [t_0, \infty)$, and the boundary condition takes the form

$$
\begin{aligned}
(3.8) \qquad \beta h_\theta(z,t) - \bar{\sigma}_\theta(z,t)\partial_z h_\theta(z,t) - \partial_t h_\theta(z,t) - \bar{\eta}_\theta(z,t) &= 0, \quad \text{on } \{\underline{z}\} \times [t_0, \infty), \\
h_\theta(z,t) &= h_\theta(z, t+T_p), \ \forall\ (z,t) \in \mathcal{U}.
\end{aligned}
$$

For semi-linear PDEs of the form (3.2), it is well known that a classical solution (i.e a solution $h_\theta(z,t)$ that is continuously differentiable) does not exist. The weak solution concept that we employ here is that of a viscosity solution (Crandall and Lions, 1983). Compared to other weak solution concepts, it allows for very general sets of boundary conditions, and also enables us to derive regularity properties of the solutions, such as Lipschitz continuity, under reasonable conditions. This is a common solution concept for equations of the HJB form; we refer to Crandall, Ishii and Lions (1992) for a user's guide, and Achdou *et al* (2017) for a useful discussion. The following ensures existence of a unique, Lipschitz continuous viscosity solution to (3.2):

Assumption 1. *(i) $G_a(x,z,t)$ and $\pi_\theta(x,z,t)$ are uniformly continuous in $(z,t) \in \mathcal{U}$, for each x, θ. Furthermore, $\bar{G}_\theta(z,t)$ and $\bar{r}_\theta(z,t)$ are Lipschitz continuous uniformly over θ.*

(ii) $\lambda(t)$ is Lipschitz continuous and bounded away from 0.

(iii) $|\lambda(t)\bar{G}_\theta(z,t)| \leq M$ for some $M < \infty$.

[14]Instead of using (3.6) as a boundary condition, we could have alternatively augmented (3.2) with (3.6), allowing for potential discontinuities in the coefficients of the PDE. This is theoretically equivalent, but the analysis of PDEs with discontinuous coefficients is rather more involved, so we do not take this route here.

(iv) $\bar{\sigma}_\theta(\underline{z}, t), \bar{\eta}_\theta(\underline{z}, t)$ are bounded and Lipschitz continuous in t uniformly over θ. Furthermore, $\bar{\sigma}_\theta(\underline{z}, t)$ is uniformly bounded away from 0, i.e $\bar{\sigma}_\theta(\underline{z}, t) \geq \delta > 0$.

The first part of Assumption 1(i) is only needed to show existence of solutions to the sample version of PDE (3.2) that will be introduced shortly. We will drop this after developing the theory further in Section 5.2. Appendix C.3 provides primitive conditions for verifying the second part of Assumption 1(i) under the soft-max policy class (2.3). In short, we shall require either the temperature parameter σ to be bounded away from 0, or that atleast one of the covariates is continuous. With purely discrete covariates and $\sigma \to 0$, $\bar{G}_\theta(z, t)$ and $\bar{r}_\theta(z, t)$ will typically be discontinuous, unless the policies depend only on x. Even with discontinuous coefficients, however, a Lipschitz continuous solution may still exist. Indeed, depending on the boundary condition, we can allow $\bar{G}_\theta(z, t), \bar{r}_\theta(z, t)$ to be discontinuous in one of the arguments, see Appendix C.3. Recent work by Barles and Chasseigne (2014) suggests just piecewise Lipschitz continuity of $\bar{G}_\theta(z, t), \bar{r}_\theta(z, t)$ may also be sufficient; we leave this extension for future research.

Assumption 1(ii) implies the arrival rates vary smoothly with t and are bounded away from 0. Assumption 1(iii) is a mild requirement ensuring changes to z are bounded. Assumption 1(iv) provides regularity conditions for the Neumann boundary condition.

Lemma 1. *Suppose that Assumptions 1 hold. Then for each θ, there exists a unique viscosity solution $h_\theta(z, t)$ to (3.2) under the boundary conditions (3.3), (3.5), (3.7) or (3.8).*

Note that (3.2) define a class of PDEs indexed by θ, the solution to each of which is the integrated value function $h_\theta(z, t)$ from following π_θ. The social planner's objective is to choose θ^* that maximizes the forecast welfare at the initial time, t_0, and initial budget, z_0:

$$\theta^* = \arg\max_{\theta \in \Theta} h_\theta(z_0, t_0). \tag{3.9}$$

The welfare criterion above presupposes that the planner only has access to a single forecast. We can alternatively allow for multiple forecasts. Denote each separate forecast for the arrival rates by $\lambda(t; \xi)$, where ξ indexes the forecasts. For example, in consensus or ensemble forecasts, each ξ may represent a different estimate or model. For each forecast ξ, we can obtain the integrated value function $h_\theta(z, t; \xi)$ by replacing $\lambda(t)$ in (3.2) with $\lambda(t; \xi)$. Let $P(\xi)$ denote some - possibly subjective - probability distribution that the social planner places over the forecasts. We take this distribution as given. Then we define the 'forecasted' integrated value function as

$$W_\theta(z, t) = \int h_\theta(z, t; \xi) dP(\xi).$$

The social planner's problem is to then choose θ^* such that

$$\theta^* = \arg\max_{\theta \in \Theta} W_\theta(z_0, t_0).$$

Our welfare criterion conditions on a forecast, or more generally, a prior over forecasts. One could alternatively calculate the welfare based on an unknown but true value of $\lambda(t)$. We analyze this alternative welfare criterion in Appendix C.4. Apart from adding an additional term to the regret, which solely depends on the estimation error of $\lambda(t)$, none of the subsequent analysis is affected. In particular, this additional term is unaffected by the complexity of the policy class.

3.1. **The sample version of the social planner's problem.** The unknown parameters in the social planner's problem are F and $r(s,a)$. As in Section 2, the social planner can leverage observational data to obtain estimates F_n and $\hat{r}(s,a)$ of F and $r(s,a)$. We discuss estimation of $\hat{r}(s,a)$ in Section (5); assume for now that a consistent estimate is available. We can then plug-in the quantities $F_n, \hat{r}(s,a)$, to obtain

$$\hat{r}_\theta(z,t) := E_{x\sim F_n}[\hat{r}(s,1)\pi_\theta(1|x,z,t)],$$

along with

$$\hat{G}_\theta(z,t) := E_{x\sim F_n}\left[G_1(x,z,t)\pi_\theta(1|x,z,t) + G_0(x,z,t)\pi_\theta(0|x,z,t)\right].$$

Based on the above we can construct the sample version of PDE (3.2) as

$$\beta\hat{h}_\theta(z,t) - \lambda(t)\hat{G}_\theta(z,t)\partial_z\hat{h}_\theta(z,t) - \partial_t\hat{h}_\theta(z,t) - \lambda(t)\hat{r}_\theta(z,t) = 0 \text{ on } \mathcal{U}, \tag{3.10}$$

together with the corresponding sample versions of the boundary conditions (3.3), (3.7) or (3.8). A unique solution to PDE (3.10) exists for each θ under Assumption 1, since, among other things, it implies $\hat{G}_\theta(z,t)$ and $\hat{r}_\theta(z,t)$ are uniformly continuous. As before, one should think of (3.10) as defining a class of PDEs indexed by θ, the solution to each of which is the integrated value function $\hat{h}_\theta(z,t)$ that can be used as an estimate for $h_\theta(z,t)$. Based on these estimates, we can now solve a sample version of the social planner's problem:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \hat{h}_\theta(z_0,t_0). \tag{3.11}$$

In the case where there are multiple forecasts, we will have $\hat{h}_\theta(z,t;\xi)$ as the solution to (3.10) for each $\lambda(t;\xi)$, and the estimated policy parameter $\hat{\theta}$ is obtained as

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \hat{W}_\theta(z_0,t_0),$$

where

$$\hat{W}_\theta(z,t) := \int \hat{h}_\theta(z,t;\xi)dP(\xi).$$

While the PDE form for $\hat{h}_\theta(z,t)$ is very convenient for our theoretical results, it is not quite useful for computing $\hat{\theta}$. So for estimation we use a discretized version of (3.10). In particular, we discretize the arrivals so that the law of motion for z is given by (here, and in what follows, we use the 'prime' notation to denote one-step ahead quantities following the current one)

$$z' = \begin{cases} z + b_n^{-1}G_a(x,z,t) & \text{if } z + b_n^{-1}G_a(x,z,t) \geq \underline{z} \\ \underline{z} & \text{otherwise} \end{cases}, \tag{3.12}$$

for some approximation factor b_n. Additionally, in the approximation scheme, the difference between arrival times is specified as

$$t' - t \sim \min\left\{\text{Exponential}(\lambda(t)b_n), T - t\right\},$$

with the truncation at T used as a device to impose a finite horizon boundary condition. To simplify the notation, we shall allow $G_a(s)$ and $r(x,1)$ to be potentially discontinuous at $z = \underline{z}$

in case of the Neunmann boundary condition, and thus avoid the need for the quantities $\bar{\sigma}_\theta(z,t)$ and $\bar{\eta}_\theta(z,t)$.[15] The rest of environment is the same as before. For this discretized setup, define $\tilde{h}_\theta(z,t)$ as the expected value function when an individual *happens* to arrive at state (z,t). This can be obtained as the fixed point to the following dynamic programming problem:

$$\tilde{h}_\theta(z,t) = \begin{cases} \frac{\hat{r}_\theta(z,t)}{b_n} + E_{n,\theta}\left[e^{-\beta(t'-t)}\tilde{h}_\theta\left(z',t'\right)|z,t\right] \\ 0 \qquad \text{for } (z,t)\in\Gamma \quad \text{(Dirichlet only)} \end{cases}. \tag{3.13}$$

Here,

$$E_{n,\theta}\left[e^{-\beta(t'-t)}\tilde{h}_\theta\left(z',t'\right)|z,t\right] := \int e^{-\beta\frac{\omega}{b_n}}E_{x\sim F_n}\left[\tilde{h}_\theta\left(z+\frac{G_1(x,t,z)}{b_n}, t+\frac{\omega}{b_n}\right)\pi(1|x,z,t)\right.$$
$$\left.+\tilde{h}_\theta\left(z+\frac{G_0(x,t,z)}{b_n}, t+\frac{\omega}{b_n}\right)\pi(0|x,z,t)\right]g_{\lambda(t)}(\omega)d\omega.$$

In particular, for any function f, $E_{n,\theta}[f(z',t')|z,t]$ denotes the joint expectation over z',t' conditional on the values of z,t and when following the policy π_θ. Precisely, the expectation is joint over three independent probability distributions: (i) The distribution F_n of the covariates, (ii) the probability distribution, $g_{\lambda(t)}(\cdot)$, over the truncated exponential waiting time process, and (iii) the probability distribution induced on z' due to the randomization of policies using $\pi_\theta(a|s)$.

We emphasize that the Neumann boundary condition does not have to be imposed explicitly since we allowed $\bar{G}_\theta$ and $\bar{r}_\theta$ to be discontinuous. The same goes for the periodic boundary condition since it would hold naturally as long as the environment is periodic in t.

The usual contraction mapping argument ensures that $\tilde{h}_\theta$ always exists as long as $T<\infty$ or $\beta<1$. We shall therefore use $\tilde{h}_\theta$ as the feasible sample counterpart of h_θ. With this, we can estimate θ^* by $\tilde{\theta}=\arg\min_{\theta\in\Theta}\tilde{h}_\theta(z_0,t_0)$. Note that the latter corresponds to solving for the optimal policy function under the sample dynamics described by $F_n,\hat{r}(s,a)$ and $\lambda(t)$. Here, both the rewards $\hat{r}(s,a)$, and the dynamics are known. This nests computation of $\hat{\theta}$ into a standard Reinforcement Learning problem for learning the optimal policy function.

3.2. **Examples.**

3.2.1. *Budget constraints.* We subsume Examples 1.1-1.3 under the common theme of budget constraints. Let z denote the current budget. Suppose that the social planner receives cash at the flow rate $\sigma(z,t)$ with respect to time, while the flow cost of treating any individual is given by $c(x,z,t)$ with respect to the expected mass m of individuals. In this case $G_a(s) = \lambda(t)^{-1}\sigma(z,t) - c(x,z,t)\mathbb{I}(a=1)$. The first term is divided by $\lambda(t)$ to measure all flows relative to the expected mass of individuals m, rather than time.

We can also consider settings with borrowing, where the rate of interest is given by b. For simplicity suppose that the borrowing rate is the same as the savings rate. Then the law of motion for z is given by

$$G_a(s) = \lambda(t)^{-1}\{\sigma(z,t)+bz\} - c(x,z,t)\mathbb{I}(a=1).$$

[15]However, we need them for the theory of viscosity solutions since it does not allow for discontinuous PDEs.

We assume b to be constant here for simplicity, but we could just as easily allow it to change with z, t. With this definition of $G_a(s)$, it is easy to see that we can use PDE (3.2) to model the behavior of $h_\theta(z, t)$ under the various constraints of finite budget or time and/or borrowing constraints, by choosing the boundary condition appropriately.

3.2.2. *Queues.* We consider the case of a single queue. Extensions to multiple queues may be found in Appendix C.2.2. The institutional variable z is now the queue length. Suppose that individuals exit the queue (i.e after they finish treatment) at some known rate $e(z, t)$ with respect to time. We may normalize the measure of z so that taking action $a = 1$ adds people to the queue at the rate 1. Then the law of motion for z is given by $\dot{z} = G_a(s)$, where $G_a(s) = \mathbb{I}(a = 1) - \lambda(t)^{-1}e(z, t)$. Note that for the environment to generate queues, we would need $e(z, t) < \lambda(t)$ for atleast some t. It is natural to setup this problem as a periodic one, with or without a nonlinear Neumann boundary condition at $z = 0$. The latter is useful if the planner would like to allow the policy to behave discontinuously between $z = 0$ (when there is no queue) and $z > 0$.

Since waiting is costly, this cost will be reflected in the flow rewards $r(s, 1)$ now being a function of z, along with x. In the simplest case, we can assume the cost is multiplicative, i.e $r(x, z, 1) = c(z)\tilde{r}(x, 1)$, where $\tilde{r}(x, 1)$ is the reward when $z = 0$, and $c(\cdot)$ is a monotonically decreasing function. We can then use observational data to estimate $\tilde{r}(x, 1)$ using doubly robust methods, while estimating $c(\cdot)$ through other means.[16] In general, however, $r(x, z, 1)$ could be non-linear in z. In such cases, we need an observational dataset that includes z or some proxy for it. The regret bound would then typically be non-parametric (see, Section 5).

3.2.3. *Capacity constraints.* Suppose that the planner faces a fixed capacity constraint. We discuss here a relatively simple version of the problem in which people are turned away if the capacity is full. One can alternatively think of augmenting this setup with queues.

The variable z now measures the amount of free capacity, assumed to take values between $[0, C]$. A value of $z = 0$ implies the capacity is full. We assume capacity is freed up, i.e people finish treatment, at the rate $e(z, t)$ with respect to time. This rate is assumed to be known or estimable. An important simplification here is that $e(\cdot, \cdot)$ does not depend on the characteristics of individuals who are currently being treated, but only on the number of people currently in the system which is $C - z$.[17] This ensures the state space is finite. As before, we normalize the measurement of capacity so that it is filled up at the rate 1 when $a = 1$. In this case $G_a(s) = \mathbb{I}(a = 1) - \lambda(t)^{-1}e(z, t)$. Thus, in this example, capacity behaves very similarly to queues, including in terms of a Neumann boundary condition at $z = 0$.[18] The main difference is that the flow rewards typically do not depend on the capacity, i.e $r(s, 1) = r(x, 1)$. Thus one can estimate them using doubly robust methods.

[16]For instance, there is a substantial literature on the effect of entering labour market programs at different times in the unemployment spell, see e.g Lechner (1999), Sianesi (2004), Crepon *et al* (2009) and Vikstrom (2017).
[17]So e.g, every individual has the same probability of moving out irrespective of how long he/she has been treated.
[18]There may additionally be another Neumann boundary condition at $z = C$ due to possible discontinuity in $e(z, t)$, see the example in Appendix C.2.1.

4. The actor-critic algorithm

In this Section we propose a Reinforcement Learning algorithm to efficiently compute $\tilde{\theta}$ in equation (3.13). For ease of exposition, we shall focus here on the Dirichlet boundary condition. Extensions to the other boundary conditions will be provided in Appendix B. For now, let us note that in the Dirichlet setting, the environment ends in some finite time. In the RL literature these are known as episodic cases. In settings where the program continues indefinitely (as with the periodic or periodic Neumann boundary conditions), our algorithms will require some modifications. One approach is to convert these environments into episodic ones by adding a finite time boundary condition $h_\theta(z,T) = 0 \ \forall z$, where T is suitably long. Alternatively, we can exploit the equivalence between discounting and random stopping to convert an infinite horizon model into an episodic one with a random horizon. We discuss these and other approaches in Appendix B, where we also characterize the numerical error involved, if any.

In a standard, episodic, Reinforcement Learning (RL) framework, an algorithm runs multiple instances, called episodes, of a dynamic environment. At any particular state, on any particular episode, the algorithm takes an action a according to the current policy function π_θ and observes the reward and the future value of the state. Based on these quantities, it updates the policy parameter to a new value θ'. The process then continues with the new updated policy function $\pi_{\theta'}$ until the parameter θ converges, or more likely, the welfare does not increase.[19]

Estimation of $\tilde{\theta}$ in equation (3.13) fits naturally in the above context, since we can simulate a 'sample' dynamic environment as follows: Suppose that the current state is $s \equiv (x,z,t)$, and the policy parameter is θ. The computer chooses an action a according to the policy function $\pi_\theta(a|s)$, which results in a reward of $\hat{r}(s,a)$. The next individual arrives at time $t' = t + \Delta t/b_n$, where $\Delta t \sim \text{Exponential}(\lambda(t))$. New values of the institutional state variable z' are obtained as in (3.12). Finally, new values of the covariates x' are drawn from the distribution $F_n(.)$, i.e each individual is drawn with replacement with probability $1/n$ from the sample set of observations. Based on the reward $\hat{r}(s,a)$ and new state $s' \equiv (x', z', t')$, the policy parameter is updated to a new value θ. This process repeats until (z,t) reach the boundary of $\mathcal{U}$. This determines the end of the current episode. Following this, we start a new episode with the starting values (z_0, t_0). We proceed in this fashion indefinitely until θ converges.

In this section, we adapt one of the most widely used RL algorithms - the Actor-Critic algorithm - to our context. We differ from the standard RL approach, however, in employing the integrated value function $\tilde{h}_\theta(z,t)$ from (3.13) as the central ingredient of our algorithm instead of the usual value function $\tilde{V}_\theta(s)$ - we explain the rationale for this in Section 4.1 below.

Actor-Critic algorithms aim to calculate $\hat{\theta}$ by updating θ at each state of each episode using stochastic gradient descent along the direction $\tilde{g}(\theta) \equiv \nabla_\theta \left[\tilde{h}_\theta(z_0,t_0)\right]$:

$$\theta \longleftarrow \theta + \alpha_\theta \tilde{g}(\theta),$$

where α_θ is the learning rate. Denote by $\tilde{Q}_\theta(s,a)$, the action-value function

$$\tilde{Q}_\theta(s,a) := \hat{r}_n(s,a) + E_{n,\theta}\left[e^{-\beta(t'-t)}\tilde{h}_\theta(z',t')|s,a\right], \tag{4.1}$$

[19]We monitor the welfare by running a test iteration of the environment periodically with the current value of θ.

where $\hat{r}_n(s,a) := \hat{r}(s,a)/b_n$, and $E_{n,\theta}[.]$ in this context denotes the expectation over the (stationary) distribution of the states s, actions a induced by the policy function π_θ in the (sample) dynamic environment of Section 3.1. The Policy-Gradient theorem (see e.g Sutton *et al*, 2000) provides an expression for $\tilde{g}(\theta)$ as

$$\tilde{g}(\theta) = E_{n,\theta}\left[e^{-\beta(t-t_0)}\tilde{Q}_\theta(s,a)\nabla_\theta \ln \pi(a|s;\theta)\right],$$

A well known result (see e.g, Sutton and Barto, 2018) is that

$$E_{n,\theta}\left[e^{-\beta(t-t_0)}\tilde{Q}_\theta(s,a)\nabla_\theta \ln \pi(a|s;\theta)\right] = E_{n,\theta}\left[e^{-\beta(t-t_0)}\left(\tilde{Q}_\theta(s,a) - b(s)\right)\nabla_\theta \ln \pi(a|s;\theta)\right]$$

for any arbitrary 'baseline' $b(.)$ that is a function of s. Let $\dot{h}_\theta(z,t)$ denote some functional approximation for $\tilde{h}_\theta(z,t)$. We exploit the fact that the continuation value of the state-action pair only depends on z,t, and therefore use $\dot{h}_\theta(z,t)$ as the baseline, which gives us

$$\tilde{g}(\theta) = E_{n,\theta}\left[e^{-\beta(t-t_0)}\left(\tilde{Q}_\theta(s,a) - \dot{h}_\theta(z,t)\right)\nabla_\theta \ln \pi(a|s;\theta)\right].$$

The above is infeasible since we don't know $\tilde{Q}_\theta(s,a)$. However we can heuristically approximate $\tilde{Q}_\theta(s,a)$ with the one step 'bootstrap' return as suggested by equation (4.1) (here the term 'bootstrap' refers to its usage in the RL literature, see Sutton and Barto, 2018):

$$R^{(1)}(s,a) = \hat{r}_n(s,a) + \mathbb{I}\{(z',t') \in \mathcal{U}\}\, e^{-\beta(t'-t)}\dot{h}_\theta(z',t'),$$

This enables us to obtain an approximation for $\tilde{g}(\theta)$ as

$$\tilde{g}(\theta) \approx E_{n,\theta}\left[e^{-\beta(t-t_0)}\delta_n(s,s',a)\nabla_\theta \ln \pi(a|s;\theta)\right], \tag{4.2}$$

where $\delta_n(s,s',a)$ is the Temporal-Difference (TD) error defined as

$$\delta_n(s,s',a) := \hat{r}_n(s,a) + \mathbb{I}\{(z',t') \in \mathcal{U}\}\, e^{-\beta(t'-t)}\dot{h}_\theta(z',t') - \dot{h}_\theta(z,t).$$

We now describe the functional approximation for $\tilde{h}_\theta(z,t)$. Let $\phi_{z,t} = (\phi_{z,t}^{(j)}, j = 1,\ldots,d_\nu)$ denote a vector of basis functions of dimension d_ν over the space of z,t. For the sake of argument, consider approximating $\tilde{h}_\theta(z,t)$ by choosing the weights ν to minimize the infeasible mean squared error criterion:

$$\arg\min_\nu \tilde{S}(\nu|\theta) \equiv \arg\min_\nu E_{n,\theta}\left[e^{-\beta(t-t_0)}\left\|\tilde{h}_\theta(z,t) - \nu^\intercal\phi_{z,t}\right\|^2\right].$$

Then we can update the value function weights, ν, using gradient descent

$$\nu \longleftarrow \nu + \alpha_\nu \nabla_\nu \tilde{S}(\nu|\theta)$$

for some value function learning rate α_ν. Here the gradient is given by

$$\tilde{\chi}(\nu|\theta) := \nabla_\nu \tilde{S}(\nu|\theta) \propto E_{n,\theta}\left[e^{-\beta(t-t_0)}\left(\tilde{h}_\theta(z,t) - \nu^\intercal\phi_{z,t}\right)\phi_{z,t}\right].$$

The above procedure is infeasible since $\tilde{h}_\theta(z,t)$ is unknown. However, as before, we can heuristically approximate $\tilde{h}_\theta(z,t)$ using the one step bootstrap return $R^{(1)}$ and obtain

$$\tilde{\chi}(\nu|\theta) \approx E_{n,\theta}\left[e^{-\beta(t-t_0)}\delta_n(s,s',a)\phi_{z,t}\right]. \tag{4.3}$$

Algorithm 1: Actor-Critic (Dirichlet boundary condition)

Initialise policy parameter weights $\theta \leftarrow 0$
Initialise value function weights $\nu \leftarrow 0$

Repeat forever:

Reset budget: $z \leftarrow z_0$
Reset time: $t \leftarrow t_0$
$I \leftarrow 1$

While $(z, t) \in \mathcal{U}$**:**

$x \sim F_n$ (Draw new covariate at random from data)

$a \sim \pi(a|s;\theta)$ (Draw action, note: $s = (x, z, t)$)

$R \leftarrow \hat{r}(s, a)/b_n$ (with $R = 0$ if $a = 0$)

$\omega \sim \text{Exponential}(\lambda(t))$

$t' \leftarrow t + \omega/b_n$

$z' \leftarrow z + G_a(x, z, t)/b_n$

$\delta \leftarrow R + \mathbb{I}\{(z', t') \in \mathcal{U}\}e^{-\beta(t'-t)}\nu^\intercal \phi_{z',t'} - \nu^\intercal \phi_{z,t}$ (Temporal-Difference error)

$\theta \leftarrow \theta + \alpha_\theta I \delta \nabla_\theta \ln \pi(a|s;\theta)$ (Update policy parameter)

$\nu \leftarrow \nu + \alpha_\nu I \delta \phi_{z,t}$ (Update value parameter)

$z \leftarrow z'$

$t \leftarrow t'$

$I \leftarrow e^{-\beta(t'-t)} I$

The heuristic for the bootstrap approximation above is based on equation (3.13), which implies that an unbiased estimator of $\tilde{h}_\theta(z, t)$ is given by sum of the current reward $\hat{r}_n(s, a)$, and the discounted future value of $\tilde{h}_\theta(z', t')$.

Using equations (4.2) and (4.3), we can now construct stochastic gradient updates for θ, ν as

$$\theta \longleftarrow \theta + \alpha_\theta e^{-\beta(t-t_0)} \delta_n(s, s', a) \nabla_\theta \ln \pi(a|s;\theta) \tag{4.4}$$

$$\nu \longleftarrow \nu + \alpha_\nu e^{-\beta(t-t_0)} \delta_n(s, s', a) \phi_{z,t}, \tag{4.5}$$

by replacing the expectations in (4.2), (4.3) with their corresponding unbiased estimates obtained from the values of state variables as they come up in each episode. Importantly, the updates (4.4) and (4.5) can be applied simultaneously on the same set of current state values, as long as $\alpha_\nu \gg \alpha_\theta$. This is an example of two-timescale stochastic gradient decent: the parameter with the lower value of the learning rate is said to be updated at the slower time scale. When the timescale for ν is much faster than that for θ, one can imagine that the value of $\nu^\intercal \phi_{z,t}$ has effectively converged to the value function estimate for current policy parameter θ. Thus we can proceed with updating θ as if its corresponding (approximate) value function were already known.

The pseudo-code for this procedure is presented in Algorithm 1.

4.1. **Basis dimensions and Integrated value functions.** The functional approximation for $\tilde{h}_\theta(z, t)$ involves choosing a vector of bases $\phi_{z,t}$ of dimension d_ν . From a statistical point of view, the optimal choice of d_ν is in fact infinity. There is no bias-variance tradeoff since we would like to compute $\tilde{h}_\theta(z,t)$ exactly. We can simply take as high a value of d_ν as computationally feasible. This useful property is a consequence of employing $\tilde{h}_\theta(z,t)$ rather than the standard value function (which is a function of x, z, t) in the Actor-Critic algorithm. Since $\hat{r}(s,a)$ could be a function of Y (as with doubly robust estimators, for example), we would need some regularization if we want to obtain a functional approximation for the standard value function, to ensure we don't overfit to the outcome data. This is not an issue for $\tilde{h}_\theta(z,t)$, however, as it only involves the expected value of $\hat{r}(s,a)$ given z, t. Thus, using $\tilde{h}_\theta(z,t)$ enables us to avoid an additional regularization term.

4.2. **Convergence of the Actor-Critic algorithm.** Our proposed algorithm differs from the standard versions of the Actor-Critic algorithm in only using the integrated value function. Consequently, its convergence follows by essentially the same arguments as that employed in the literature for actor-critic methods, see e.g, Bhatnagar *et al* (2009). In this section, we restate their main results, specialized to our context. Since all of the convergence proofs in the literature are obtained for discrete Markov states, we need to impose the technical device of discretizing time and making it bounded, so that the states are now discrete (the other terms z and x are already discrete, the latter since we use empirical data). This greatly simplifies the convergence analysis, but does not appear to be needed in practice.

Let $\mathcal{S}$ denote the set of all possible values of (z,t), after discretization. Also, denote by Φ, the $|\mathcal{S}| \times d_\nu$ matrix whose ith column is $(\phi_{z,t}^{(i)}, (z,t) \in \mathcal{S})^\intercal$, where $\phi_{z,t}^{(i)}$ is the ith element of $\phi_{z,t}$.

Assumption C. *(i) $\pi_\theta(a|s)$ is continuously differentiable in θ for all s, a.*

(ii) The basis functions $\{\phi_{z,t}^{(i)} : i : 1, \ldots, d_\nu\}$ are linearly independent, i.e Φ has full rank. Also, for any vector ν, $\Phi\nu \neq e$, where e is the $\mathcal{S}$-dimensional vector with all entries equal to one.

(iii) The learning rates satisfy $\sum_k \alpha_\nu^{(k)} \to \infty$, $\sum_k \alpha_\nu^{(k)2} < \infty$, $\sum_k \alpha_\theta^{(k)} \to \infty$, $\sum_k \alpha_\theta^{(k)2} < \infty$ and $\alpha_\theta^{(k)}/\alpha_\nu^{(k)} \to 0$ where $\alpha_\theta^{(k)}, \alpha_\nu^{(k)}$ denote the learning rates after k steps/updates of the algorithm.

(iv) The update for θ is bounded i.e

$$\theta \longleftarrow \Gamma\left(\theta + \alpha_\theta \delta_n(s, s', a) \nabla_\theta \ln \pi(a|s;\theta)\right)$$

where $\Gamma : \mathbb{R}^{dim(\theta)} \to \mathbb{R}^{dim(\theta)}$ is a projection operator such that $\Gamma(x) = x$ for $x \in C$ and $\Gamma(x) \in C$ for $x \notin C$, where C is any compact hyper-rectangle in $\mathbb{R}^{dim(\theta)}$.

Differentiability of π_θ with respect to θ is a minimal requirement for all Actor-Critic methods. Assumption C(ii) is also mild and rules out multicollinearity in the basis functions for the value approximation. Assumption C(iii) places conditions on learning rates that are standard in the literature of stochastic gradient descent with two timescales. Assumption C(iv) is a technical condition imposing boundedness of the updates for θ. This is an often used technique in the analysis of stochastic gradient descent algorithms. Typically this is not needed in practice, though it may sometimes be useful to bound the updates when there are outliers in the data.

Define $\mathcal{Z}$ as the set of local maxima of $J(\theta) \equiv \tilde{h}_\theta(z_0, t_0)$, and $\mathcal{Z}^\epsilon$ an ϵ-expansion of that set. Also, $\theta^{(k)}$ denotes the k-th update of θ. We then have the following theorem on the convergence of our Actor-Critic algorithm.

Theorem 1. *Suppose that Assumptions C hold and additionally that $\nabla_\theta \pi_\theta(s)$ is uniformly Hölder continuous in s. Then, for each $\epsilon > 0$, there exists M such that if $d_\nu \geq M$, then $\theta^{(k)} \to \mathcal{Z}^\epsilon$ with probability 1 as $k \to \infty$.*

The above theorem is for the most part a direct consequence of the results of Bhatnagar *et al* (2009). We provide further discussion and a justification of the result in Appendix B. For the exponential soft-max functional as in (2.3), $\mathcal{Z}$ is actually a singleton under discrete states (Thomas, 2014). So our algorithm will converge to the global optimum.

4.3. **Multiple forecasts.** Thus far, we have only considered the case with a single forecast. The extension to multiple forecasts is straightforward: we simply draw a value of ξ from $P(\xi)$ at the start of every new episode. In consensus or ensemble forecasts this just means drawing an estimate or model at random based on the weights given to each of them. In other cases, e.g if $\lambda(t)$ follows a continuous time AR(1) process $d\lambda(t) = -\phi\lambda(t)dt + \sigma dB_t$ (where dB_t denotes the increments to standard Brownian motion), we would draw the increments at random from a normal distribution before each update step, but the cumulative effect of this is equivalent to drawing an infinite dimensional parameter ξ at the beginning of each episode.

4.4. **Parallel updates.** While Theorem 1 assures convergence of our algorithm, in practice the updates could be volatile and may take a long time to converge. Much of the reason for this is the correlation between the updates as one cycles through each episode - indeed, note that the state pairs (s, s') are highly correlated. Hence the stochastic gradients become correlated and one needs many episodes to move in the direction of the true (i.e the expected) gradient. This is a common problem for all Actor-Critic algorithms, but recently Mnih *et al* (2015) have proposed to solve this through the use of asynchronous parallel updates. The idea is to run multiple versions of the dynamic environment on parallel threads or processes, each of which independently and asynchronously updates the shared global parameters θ and v. Since at any given point in time, the parallel threads are at a different point in the dynamic environment (they are started with slight offsets), successive updates are decorrelated. Additionally, the algorithm is faster by dint of being run in parallel.

Algorithm 2 provides the pseudo-code for parallel updating. It also amends the previous version of the algorithm by adding batch updates. In batch updating, the researcher chooses a batch size B such that the parameter updates occur only after averaging over B observations. This usually results in a smoother update trajectory because extreme values of the updates are averaged out.

4.5. **Choosing the tuning parameters.** To implement our algorithm, we need to specify the basis functions for the value approximation and the learning rates. In specifying these, one should try incorporate prior knowledge about the environment. Indeed, the ability to do so is one of the strengths of the algorithm. For instance, in the time constrained boundary condition, we know that $\tilde{h}_\theta(z, 0) = 0$ for all z. So in this setting, the basis functions could be chosen so that there are also 0 when $t = 0$. In a similar vein, in the periodic boundary condition setting, one could choose the bases so that they are also periodic in t.

Algorithm 2: Parallel Actor-Critic (Dirichlet boundary condition)

Initialise policy parameter weights $\theta \leftarrow 0$
Initialise value function weights $\nu \leftarrow 0$
Batch size B

For $p = 1, 2, ...$ processes, launched in parallel, each using and updating the same global parameters θ and ν:

Repeat forever:

Reset budget: $z \leftarrow z_0$
Reset time: $t \leftarrow t_0$
$I \leftarrow 1$

While $(z, t) \in \mathcal{U}$ **:**

$\theta_p \leftarrow \theta$ (Create local copy of θ for process p)

$\nu_p \leftarrow \nu$ (Create local copy of ν for process p)

batch_policy_upates $\leftarrow 0$

batch_value_upates $\leftarrow 0$

For $b = 1, 2, ..., B$**:**

$x \sim F_n$ (Draw new covariate at random from data)

$a \sim \pi(a|s; \theta_p)$ (Draw action, note: $s = (x, z, t)$)

$R \leftarrow \hat{r}(s, a)/b_n$ (with $R = 0$ if $a = 0$)

$\omega \sim \text{Exponential}(\lambda(t))$

$t' \leftarrow t + \omega/b_n$

$z' \leftarrow z + G_a(x, z, t)/b_n$

$\delta \leftarrow R + \mathbb{I}\{(z', t') \in \mathcal{U}\}e^{-\beta(t'-t)}\nu_p^\intercal \phi_{z',t'} - \nu_p^\intercal \phi_{z,t}$ (TD error)

batch_policy_upates $\leftarrow$ batch_policy_upates $+ \alpha_\theta I \delta \nabla_\theta \ln \pi(a|s; \theta_p)$

batch_value_upates $\leftarrow$ batch_value_upates $+ \alpha_\nu I \delta \phi_{z,t}$

$z \leftarrow z'$

$t \leftarrow t'$

$I \leftarrow e^{-\beta(t'-t)} I$

If $(z, t) \notin \mathcal{U}$, break **For**

Globally update: $\nu \leftarrow \nu +$ batch_value_upates$/B$
Globally update: $\theta \leftarrow \theta +$ batch_policy_upates$/B$

Choosing the right learning rates may require some experimentation. These are typically taken to be constant, rather than decaying over time as the theory requires. In practice, as long as they are set small enough, this only means that the parameters will oscillate a bit around their optimum values. A common rule of thumb (see, e.g, Sutton and Barto, 2018) is to set $\alpha_\nu = 0.1/E_{n,\theta}\left[||\phi_{z,t}||\right]$, while keeping $\alpha_\theta = 0$ in the beginning. Once learning is stable, the value of α_θ can be increased slowly. There are now automated procedures for determining the rates based on Population Based Training, see e.g Jaderberg *et al* (2018). This could be an attractive choice in problems with very large state spaces.

5. Statistical and numerical properties

In this section, we analyze the statistical and numerical properties of the estimated welfare maximizing policy functions. The main result of this section is a probabilistic bound on the regret defined as the maximal difference between the integrated value functions $h_{\hat{\theta}}(z_0, t_0)$ and $h_{\theta^*}(z_0, t_0)$. We derive this using our bound on the maximal difference in the value functions

$$\sup_{(z,t)\in\bar{\mathcal{U}},\theta\in\Theta} |\hat{h}_\theta(z,t) - h_\theta(z,t)| \tag{5.1}$$

since

$$h_{\theta^*}(z_0, t_0) - h_{\hat{\theta}}(z_0, t_0) \leq 2 \sup_{(z,t)\in\bar{\mathcal{U}},\theta\in\Theta} |\hat{h}_\theta(z,t) - h_\theta(z,t)|.$$

We maintain Assumption 1 that is required for the existence of the value functions. In addition, we impose the following:

Assumption 2. *(i) (Bounded rewards) There exists $M < \infty$ such that $|Y(0)|, |Y(1)| \leq M$.*

(ii) In the Dirichlet setting, if $\underline{z} > -\infty$ in (3.4), then $\beta > 0$ and there exists $\delta > 0$ such that $\bar{G}_\theta(z,t) < -\delta$.

(iii) (Complexity of the policy function space) The collection of functions

$$\mathcal{I} = \left\{\pi_\theta(1|\cdot, z, t) : (z,t) \in \bar{\mathcal{U}}, \theta \in \Theta\right\}$$

over the covariates x, indexed by z, t and θ, is a VC-subgraph class with finite VC index v_1. Furthermore, for each $a = 0, 1$, the collection of functions

$$\mathcal{G}_a = \left\{\pi_\theta(a|\cdot, z, t)G_a(\cdot, z, t) : (z,t) \in \bar{\mathcal{U}}, \theta \in \Theta\right\}$$

over the covariates x is also a VC-subgraph class with finite VC index v_2. We shall let $v = \max\{v_1, v_2\}$.

Assumption 2(i) ensures that the rewards are bounded. This is a common assumption in the treatment effect literature (see e.g Kitagawa and Tetenov, 2018) and imposed mainly for ease of deriving the theoretical results.

Assumption 2(ii) is required only in the Dirichlet setting, and even here, only where the boundary condition is determined fully or in part by z.[20] The assumption ensures $h_\theta(z,t)$ is continuous near the boundary, by ruling out cases where the solution diverges widely for small differences in z.[21] In the subset of our examples where this assumption is relevant, i.e Examples 1.1 & 1.3, it can be verified that $\bar{G}_\theta(z,t) < 0$ (e.g, the budget can only be depleted). In such cases, Assumption (ii) restricts the policy function class to ensure there is always some expected change to the budget at any given state. This is a mild restriction: as long as there exist some people that benefit from treatment and $\beta \neq 0$, it is a dominant strategy to treat at least some fraction of the population.

[20]We can relax this assumption to: $\bar{G}_\theta(z,t) < -\delta$ on $\mathcal{N}^\epsilon$, where $\mathcal{N}^\epsilon := \{(z,t) \in \mathcal{U} : |z - \underline{z}| < \epsilon\}$ for some $\epsilon > 0$. Since it does not materially affect our examples, we use the stronger version for simplicity.

[21]E.g, if $G_\theta(z_1, t_1) = 0$ at some (z_1, t_1) close to the boundary, then as time proceeds, it is possible for z to move away from the boundary for some starting values in a neighborhood of (z_1, t_1), while moving towards the boundary for other values.

Assumption 2(iii) has already been discussed in some detail in Section 2. In many of the examples we consider, $G_a(s)$ is independent of x, as in equation (3.1). In this case it is easy to verify that $v_1 = v_2$. Let us also point out that the domain of (z,t) is $\bar{\mathcal{U}}$. In settings with Dirichlet and Neumann boundary conditions, this means that $\mathcal{I}$ and $\mathcal{G}_a$ are defined by continuously extending $\pi_\theta(1|\cdot)$ and $G_a(\cdot)$ to the boundary using the limit operation, even though the actual 'policy' and law of motion at the boundary may be quite different.

The next set of assumptions relate to the properties of the observational or RCT dataset from which we estimate $\hat{r}(s,a)$. For now we shall focus on the situation where neither time t, nor the institutional variable z affect the utilities $Y(a)$. Thus $r(s,a) \equiv r(x,a)$. Under this setting we can use doubly robust estimates, $\hat{r}(x,a)$, of the rewards to obtain a parametric bound on the regret. When z,t are able to affect $Y(a)$, we are not aware of any doubly robust estimate for the rewards. In this case, the regret will only converge to 0 at non-parametric rates. The characterization of the regret in this more general case will be provided in the next sub-section.

Assumption 3. *(i) (iid draws from F) The observed data is an iid draw of size n from the distribution F.*

(ii) (Selection on observables) $(Y(1), Y(0)) \perp W|X$.

(iii) (Strict overlap) There exists $\kappa > 0$ such that $p(x) \in [\kappa, 1-\kappa]$ for all x.

(iv) $E[Y(a)|s] = E[Y(a)|x]$ i.e, the individual outcomes do not depend on z,t.

Assumption 3(i) assumes that the observed data is representative of the entire population. If the observed population only differs from F in terms of the distribution of some observed covariates, we can reweigh the rewards, and our theoretical results continue to apply. Assumption 3(ii) assumes that the observed data is taken from an observational study that satisfies unconfoundedness. In Section 6.1, we consider extensions to non-compliance. Assumption 3(iii) ensures that the propensity scores are strictly bounded away from 0 and 1. Both Assumptions 3(ii) and 3(iii) are directly satisfied in the case of RCT data. As noted earlier, assumption 3(iv) will be relaxed in the next sub-section.

Under Assumptions 2 and 3, one can propose many different estimates of the rewards $\hat{r}(x,1)$ that are consistent for $r(x,1)$. In this paper we recommend doubly robust estimates. This is given by

$$\hat{r}(x,1) = \hat{\mu}(x,1) - \hat{\mu}(x,0) + (2W-1)\frac{Y - \hat{\mu}(x,W)}{W\hat{p}(x) + (1-W)(1-\hat{p}(x))}, \tag{5.2}$$

where $\hat{\mu}(x,w)$ and $\hat{p}(x)$ are non-parametric estimates of $\mu(x,w)$ and $p(x)$. To simplify matters, we shall assume that these non-parametric estimates are obtained through cross-fitting (Chernozhukov *et al*, 2018). This is done as follows: We divide the data randomly divided into K folds of equal size, and for each fold j, we run a machine learning estimator of our choice on the other $K-1$ folds to estimate $\hat{\mu}^{(-j)}(x,w)$ and $\hat{p}^{(-j)}(x)$. Then for any observation x_j in some fold j, we set $\hat{\mu}(x_j,w) = \hat{\mu}^{(-j)}(x_j,w)$ and $\hat{p}(x_j) = \hat{p}^{(-j)}(x_j)$. We employ cross-fitting estimators as they require minimal assumptions. Additionally, they have excellent bias properties as demonstrated by Chernozhukov *et al*, (2018), and Athey and Wager (2018). We impose the following high level conditions for the machine learning methods used in our cross-fitted estimates:

Assumption 4. *(i) There exists an $a > 0$ such that for $w = 0, 1$*

$$\sup_x |\hat{\mu}(x,w) - \mu(x,w)| = O_p(n^{-a}), \quad \sup_x |\hat{p}(x) - p(x)| = O_p(n^{-a}).$$

(ii) (L^2 convergence) There exists some $\xi > 1/2$ such that

$$E\left[|\hat{\mu}(x,w) - \mu(x,w)|^2\right] \lesssim n^{-\xi}, \qquad E\left[|\hat{p}(x) - p(x)|^2\right] \lesssim n^{-\xi}.$$

Assumption 4 is taken from Athey and Wager (2018). The requirements imposed are weak and satisfied by almost all non-parametric estimators including series regression or LASSO. Using Assumptions 1-4, one can show that the quantities $\hat{r}_\theta(z,t), \hat{G}_\theta(z,t)$ are uniformly close to $\bar{r}_\theta(z,t), \bar{G}_\theta(z,t)$. In particular, there exists a universal constant C_0 such that with probability approaching 1,

(5.3)
$$\sup_{(z,t)\in\bar{\mathcal{U}},\theta\in\Theta} |\hat{r}_\theta(z,t) - \bar{r}_\theta(z,t)| \leq C_0\sqrt{\frac{v_1}{n}} \text{ and}$$
$$\sup_{(z,t)\in\bar{\mathcal{U}},\theta\in\Theta} \left|\hat{G}_\theta(z,t) - \bar{G}_\theta(z,t)\right| \leq C_0\sqrt{\frac{v_2}{n}}.$$

The above inequalities are based on Athey and Wager (2018).[22]

From (5.3), we find that the parameters characterizing the PDEs (3.2) and (3.10) are uniformly close. This suggests that the solutions to these PDEs should also be uniformly close. The intuition is formalized in the following theorem:

Theorem 2. *Suppose that Assumptions 1-4 hold. Then with probability approaching one under the distribution F,*

$$\sup_{(z,t)\in\bar{\mathcal{U}},\theta\in\Theta} \left|\hat{h}_\theta(z,t) - h_\theta(z,t)\right| \leq C\sqrt{\frac{v}{n}},$$

under the boundary conditions (3.3) and (3.7). Furthermore, there exists $\beta_0 > 0$ that depends only on the upper bounds for $\lambda(t)$ and $\bar{G}_\theta(\cdot)$ such that the above result also holds true under the boundary conditions (3.5) and (3.8) as long as the discount factor $\beta \geq \beta_0$.

A consequence of the above is

$$h_{\theta^*}(z_0,t_0) - h_{\hat{\theta}}(z_0,t_0) \leq 2C\sqrt{\frac{v}{n}}.$$

The above statements hold uniformly over all F if similarly uniform versions of Assumptions 1-4 hold.

Theorem 2 requires the discount factor β to be sufficiently large in infinite horizon settings. We emphasize that β can be arbitrary (and even potentially negative) in finite horizon settings, such as the ones implied by the boundary conditions (3.3) and (3.7). To see what could go wrong with an infinite horizon, consider Example 1.2 with a constant flow of income and arrival rates that are independent of t. This is equivalent to setting $\partial_t h_\theta = 0$ in PDE (3.2). Suppose now $\beta \approx 0$. This implies $\partial_z h_\theta \approx \bar{r}_\theta(z)/\bar{G}_\theta(z)$. But $\bar{G}_\theta(z)$ has to be 0 at some value of z in this example (otherwise the budget will continuously increase or decrease), even as $\bar{r}_\theta(z)$ may be

[22]If the propensity score is known, one can use inverse probability weighting as in Kitagawa and Tetenov (2018) instead of the doubly robust estimate. In this case $\hat{r}(x,1) = WY/p(x) - (1-W)Y/(1-p(x))$.

non-zero. At these points h_θ varies too rapidly for $\hat{h}_\theta$ to approximate it at any reasonable rate. Hence we need β to be sufficiently large to prevent this. Indeed, this requirement is standard for analyzing viscosity solutions in infinite horizon settings, see e.g Crandall and Lions (1983), Souganidis (1985) and Barles & Lions (1991).

It seems likely that the $\sqrt{n}$ rate for the regret $h_{\theta^*}(z_0, t_0) - h_{\hat{\theta}}(z_0, t_0)$ cannot be improved upon, especially since Kitagawa and Tetenov (2018) show that this rate is optimal in the static case. However, we do not know if the VC dimension v in the rate is necessarily tight. But it does allow us to bound the difference between the integrated value functions at other values of (z, t) apart from (z_0, t_0). Indeed, Theorem 2 also implies

$$\sup_{(z,t)\in\mathcal{H}_n} \left|h_{\theta^*}(z, t) - h_{\hat{\theta}}(z, t)\right| \le 2C\sqrt{\frac{v}{n}}, \quad \text{where } \mathcal{H}_n \equiv \left\{(z, t) : \hat{h}_{\hat{\theta}}(z, t) \ge \hat{h}_{\theta^*}(z, t)\right\}.$$

We now discuss how the results extend to multiple forecasts. Assume that Assumption 1 now holds uniformly in ξ i.e, $\lambda(t; \xi)$ is bounded and Lipschitz continuous, both uniformly in ξ. Then a straightforward modification of the proof of Theorem 2 implies

$$\sup_{\xi} \sup_{(z,t)\in\bar{\mathcal{U}},\theta\in\Theta} \left|\hat{h}_\theta(z, t; \xi) - h_\theta(z, t; \xi)\right| \le C\sqrt{\frac{v}{n}}.$$

We thus have the following corollary:

Corollary 1. *Suppose that Assumptions 1-4 hold, with Assumption 1 holding uniformly in $\lambda(t; \xi)$ for all ξ. Then with probability approaching one under joint the probability distribution $F \times P(\xi)$,*

$$\sup_{(z,t)\in\bar{\mathcal{U}},\theta\in\Theta} \left|\hat{W}_\theta(z, t) - W_\theta(z, t)\right| \le C\sqrt{\frac{v}{n}}.$$

The above result holds under the boundary conditions (3.3) & (3.7) for all $\beta \in \mathbb{R}$, and also under (3.5) & (3.8) for all $\beta \ge \beta_0$. In particular, we also have

$$W_{\theta^*}(z_0, t_0) - W_{\hat{\theta}}(z_0, t_0) \le 2C\sqrt{\frac{v}{n}}.$$

5.1. **Regret bounds when the utilities are affected by z, t.** Here we describe how our results could be modified if z, t is able to affect the individual utilities $Y(a)$. Such a situation occurs in the example with queues (Example 1.4). Here z denotes the queue length, or waiting time. Since waiting is costly, the effect of this should be reflected in the outcome $Y(a)$ of the individual, so that now $E[Y(a)|s] = \mu_a(x, z)$. We assume that consistent estimation of $\mu_a(x, z)$ is possible as described in Section 3.2.2. More generally, $\mu_a(\cdot)$ may also vary with time, e.g if the cost of treatment is different in different times of the year. Typically, the effect of t is either known deterministically (as with the variable cost of treatment), or can be estimated from the observational data if the latter also includes a time variable. In all these cases, we can construct a non-parametric estimate $\hat{\mu}_a(s)$ of $\mu_a(s)$. Following this, we can estimate the rewards as

$$\hat{r}(s, 1) = \hat{\mu}_1(s) - \hat{\mu}_0(s).$$

The rest of the quantities are obtained as usual, e.g $\bar{r}_\theta(z, t) := E[\hat{r}(s, 1)\pi_\theta(1|z, t)]$ etc.

Suppose that there exists a sequence ψ_n such that, for $a \in \{0, 1\}$,

$$\sup_{x,(z,t)\in\bar{\mathcal{U}}} |\hat{\mu}_a(x, z, t) - \mu_a(x, z, t)| = O_p(\psi_n^{-1}). \tag{5.4}$$

Primitive conditions for the above can be obtained on a case-by-case basis. Also, letting VC$(\cdot)$ denote the VC dimension, suppose that for $a \in \{0, 1\}$,

$$\text{VC}\left(\bar{\mathcal{I}}_a\right) < \infty; \text{ where } \bar{\mathcal{I}}_a := \left\{\mu_a(\cdot, z, t)\pi_\theta(1|\cdot, z, t) : (z, t) \in \bar{\mathcal{U}}, \theta \in \Theta\right\}. \tag{5.5}$$

Under these two assumptions, following the analysis of Kitagawa and Tetenov (2018, Theorem 2.5), we can show[23]

$$\sup_{(z,t)\in\bar{\mathcal{U}},\theta\in\Theta} |\hat{r}_\theta(z, t) - \bar{r}_\theta(z, t)| = O_p(\psi_n^{-1}).$$

We thus have the following counterpart to Theorem 2:

Corollary 2. *Suppose that Assumptions 1-3 hold, along with (5.4) and (5.5). Then with probability approaching one under F,*

$$h_{\theta^*}(z_0, t_0) - h_{\hat{\theta}}(z_0, t_0) \le C\psi_n^{-1}$$

for some $C < \infty$. The above result holds under the boundary conditions (3.3) & (3.7) for all $\beta \in \mathbb{R}$, and also under (3.5) & (3.8) for all $\beta \ge \beta_0$.

The proof follows by the same reasoning as that for Theorem 2, and is omitted.

5.2. **Approximation and numerical convergence.** In Section 3.1, we pointed out that for computation, it is preferable to use an approximate version of PDE (3.10), given by (3.13). Our algorithm in Section 4 was based on this. Implementing this algorithm requires choosing a 'approximation' factor b_n. Here we characterize the numerical error resulting from any particular choice of b_n. This is the PDE counterpart of the analysis in Section 2.2.

For each $\theta \in \Theta$, denote by $\tilde{h}_\theta(z, t)$ the solution to (3.13),.

Theorem 3. *Suppose that Assumptions 1-4 hold and $E_{x\sim F}\left[|G_a(x, z, t)|^2\right] \le C < \infty$ for all $a \in \{0, 1\}$ and $(z, t) \in \bar{\mathcal{U}}$. Then, with probability approaching one under F, there exists $K < \infty$ independent of θ, z, t such that*

$$\sup_{(z,t)\in\bar{\mathcal{U}},\theta\in\Theta} \left|h_\theta(z, t) - \tilde{h}_\theta(z, t)\right| \le K\left(\sqrt{\frac{v}{n}} + \sqrt{\frac{1}{b_n}}\right).$$

The above result holds under the boundary conditions (3.3) & (3.5).

From the proof, we note that Theorem 3 still holds if we drop the first part of Assumption 1(i), which is only needed for results involving the sample PDE (3.10). We conjecture that Theorem 3 also holds for the Neumann boundary conditions though we were unable to prove this with our current techniques. By the results of Barles and Souganidis (1991), it is actually straightforward to show point-wise convergence of $\tilde{h}_\theta$ to h_θ for each θ, under all the boundary conditions. However, their results do not appear to be powerful enough to show uniform (over

[23]On the other hand, the rate for $\left|\hat{G}_\theta(z, t) - \bar{G}_\theta(z, t)\right|$ in the second part of (5.3) is unaffected.

θ) convergence, or to get a bound on the approximation error as we do here. In the theorem, the approximation error is given by the $b_n^{-1/2}$ term. This bound appears to be sharp under our assumptions, see, e.g, Krylov (2005). The bound is of a smaller order than b_n^{-1} obtained in Section 2.2 for ODEs. One can understand the difference in the rates as the price for dealing with viscosity solutions that are not differentiable everywhere.

Let $\tilde{\theta}$ denote the numerical approximation to $\hat{\theta}$, obtained as the solution to

$$\tilde{\theta} = \arg\max_{\theta \in \Theta} \tilde{h}_\theta(z_0, t_0).$$

A direct consequence of 3 is that

$$h_{\theta^*}(z_0, t_0) - h_{\tilde{\theta}}(z_0, t_0) \leq 2K\sqrt{\frac{v}{n}} + 2K\sqrt{\frac{1}{b_n}}.$$

Hence, as a rule of thumb, we recommend setting b_n to be some multiple of, or exactly equal to n, so that the approximation is of the same, or smaller order than the regret rates.

6. Extensions

6.1. **Non-compliance.** A common issue with observational data is that there is substantial non-compliance. Here we show how our methods can be modified to account for this. For ease of exposition, we shall specialize to examples with a budget constraint (Examples 1.1-1.3), and also suppose that the rewards are independent of z, t. We shall also assume that the treatment assignment behaves similarly to a monotone instrumental variable in that we can partition individuals into three categories: compliers, always-takers and never-takers.

We will suppose that the social planner cannot change the compliance behavior of any individual. Then the only category of people for whom a social planner can affect a welfare change are the compliers. As for the always-takers and never-takers, the planner has no control over their choices, so its equivalent to assume that the planner would always treat the former and never treat the latter. Formally, the change in reward (conditional on the covariates) for the social planner from treating an individual i, as compared to not treating, is

$$r(x_i, 1) = \begin{cases} \text{LATE}(x_i) & \text{if } i \text{ is a complier} \\ 0 & \text{otherwise,} \end{cases} \tag{6.1}$$

where $\text{LATE}(x)$ denotes the local average treatment effect for an individual with covariate x. As before, we normalize $r(x, 0)$ to 0 as we only consider expected welfare. Note that always takers and never-takers are associated with 0 rewards. The evolution of the budget is also different for each group. In particular,

$$N(z' - z) = \begin{cases} G_a(x, t, z) & \text{if } i \text{ is a complier} \\ G_1(x, t, z) & \text{if } i \text{ is an always-taker} \\ G_0(x, t, z) & \text{if } i \text{ is a never-taker.} \end{cases} \tag{6.2}$$

While the planner does not know the true compliance behavior of any individual, she can form expectations over them given the observed covariates. Let $q_c(x), q_a(x)$ and $q_n(x)$ denote the probabilities that an individual is respectively a complier, always-taker or never-taker conditional on x. Given these quantities, the analysis under non-compliance proceeds analogously to Section 3. In particular, let $h_\theta(z,t)$ denote the integrated value function in the current setting. Then we have the following PDE for the evolution of $h_\theta(z,t)$:

$$\beta h_\theta(z,t) - \lambda(t)\bar{G}_\theta(z,t)\partial_z h_\theta(z,t) - \partial_t h_\theta(z,t) - \lambda(t)\bar{r}_\theta(z,t) = 0, \quad \text{on } \mathcal{U},$$

together with the relevant boundary conditions from (3.3), (3.5), (3.7) or (3.8), where

$$\bar{r}_\theta(z,t) := E_{x\sim F}\left[q_c(x)\pi_\theta(1|x,z,t)r(x,1)\right],$$

and (in view of equation 6.2),

$$\begin{aligned}\bar{G}_\theta(z,t) := E_{x\sim F}\big[&q_c(x)\left\{\pi_\theta(1|z,t)G_1(x,t,z) + \pi_\theta(0|z,t)G_0(x,t,z)\right\} \\ &+ q_a(x)G_1(x,t,z) + q_n(x)G_0(x,t,z)\big].\end{aligned}$$

In order to estimate the optimal policy rule, we need estimates of $q_c(x), q_a(x), q_n(x)$, along with LATE(x). To obtain these, we assume that the planner has access to an observational study involving Z as the intended treatment status or instrumental variable, and W as the observed treatment. As before, Y is the observed outcome variable. Observe that $q_a(x) = E[W|X=x, Z=0]$ and $q_n(x) = E[1-W|X=x, Z=1]$. Hence we can estimate $\hat{q}_a(x)$ by running a Logit regression of W on X for the sub-group of the data with $Z=0$. Estimation of $\hat{q}_n(x)$ can be done in an analogous manner. Using both these estimates, we can also obtain $\hat{q}_c(x) = 1-\hat{q}_a(x)-\hat{q}_n(x)$. To estimate LATE$(x)$, we recommend the doubly robust version of Belloni *et al* (2017), denoted by $\widehat{\text{LATE}}(x)$. In the case where there are no always-takers, the expression for this simplifies and is given by

$$\widehat{\text{LATE}}(x) = \theta_y(1) - \theta_y(0),$$

where

$$\theta_y(1) := \frac{\hat{E}[WY|x,Z=1] + \frac{Z}{\hat{p}(x)}(WY - \hat{E}[WY|x,Z=1])}{\hat{q}_c(x) + \frac{Z}{\hat{p}(x)}(W - \hat{q}_c(x))}, \qquad \text{and}$$

$$\theta_y(0) := \frac{\hat{E}[(1-W)Y|x,Z=1] + \frac{Z}{\hat{p}(x)}\left[(1-W)Y - \hat{E}[(1-W)Y|x,Z=1]\right] - \left[\hat{\mu}(x,0) + \frac{1-Z}{1-\hat{p}(x)}(Y - \hat{\mu}(x,0))\right]}{\frac{Z}{\hat{p}(x)}\left(\hat{q}_c(x) - W\right) - \hat{q}_c(x)}.$$

In these equations, $\hat{E}[\cdot|x,Z=1]$ is an estimator for $E[\cdot|x,Z=1]$, which can be obtained through series regression, or other non-parametric methods. Additionally $\hat{p}(x)$ is an estimator for $p(x) = P(Z=1|X=x)$ - the IV propensity score.

Given the estimates $\hat{q}_c(x), \hat{q}_a(x), \hat{q}_n(x)$ and $\widehat{\text{LATE}}(x)$, it is straightforward to modify the algorithm in Section 4 to allow for non-compliance. The main difference from Algorithm 2 is that at each update we would randomly draw the compliance nature of the individual from a multinomial distribution with probabilities $(\hat{q}_c(x), \hat{q}_a(x), \hat{q}_n(x))$. Conditional on this draw, the rewards are given by sample counterpart of (6.1), and the updates to budget by (6.2). The pseudo-code for the resulting algorithm is provided in Appendix B.

Probabilistic bounds on the regret for the estimated policy rule can also be obtained by the same techniques as in Section 5. If $q_c(x), q_a(x), q_n(x)$ were known exactly, it is possible to show that the rates for the regret remain unchanged at $\sqrt{v/n}$. A similar analysis using the estimated quantities $\hat{q}_c(x), \hat{q}_a(x), \hat{q}_n(x)$ is however more involved; we leave the details for future research.

6.2. **Time varying distribution of covariates.** In realistic settings, different individuals not only respond differently to treatment, but also have (potentially) different dynamics regarding their arrival rates. Thus the distribution F_t of the covariates may change with time (and in general be different from F, the limit of the empirical distribution F_n). Assuming that the support of $F_t(\cdot)$ lies within that of $F(\cdot)$ for all t, we can write

$$F_t(y) = \int_{x \leq y} w_t(x) dF(x),$$

for some weight function $w_t(\cdot)$. Let $\lambda_x(t)$ denote the covariate specific arrival process, and $f(\cdot)$ the density function of $F(\cdot)$. Then we can note that

$$w_t(x) = \frac{\lambda_x(t)/f(x)}{\int (\lambda_\omega(t)/f(\omega)) dF(\omega)}.$$

Our previous results amounted to assuming $\lambda_x(t) = f(x)$.

With the above in mind, the PDE for the evolution of $h_\theta(z,t)$ is the same as (3.2), but now

$$\begin{aligned}
\bar{r}_\theta(z,t) &:= E_{x\sim F_t}\left[\pi_\theta(1|x,z,t) r(x,1)\right], \\
\bar{G}_\theta(z,t) &:= E_{x\sim F_t}\left[G_1(x,z,t)\pi_\theta(1|x,z,t) + G_0(x,z,t)\pi_\theta(0|x,z,t)\right]
\end{aligned}$$

and $\lambda(t)$ is replaced by $\bar{\lambda}(t)$, where

$$\bar{\lambda}(t) := \int \lambda_\omega(t) d\omega.$$

We have assumed here for simplicity that the rewards do not depend on (z,t). If the weight function $w_t(x)$ were known, we can replace F_t with its empirical counterpart $F_{n,t} := n^{-1}\sum_i w_t(x_i)\delta(x_i)$, where $\delta(\cdot)$ denotes the Dirac delta function. Thus $F_{n,t}$ is akin to a weighted empirical distribution. We can then construct the empirical PDE (3.10) using the sample quantities

$$\begin{aligned}
\hat{r}_\theta(z,t) &= E_{x\sim F_{n,t}}\left[\pi_\theta(1|x,z,t)\hat{r}(x,1)\right], \text{ and} \\
\hat{G}_\theta(z,t) &= E_{x\sim F_{n,t}}\left[G_1(x,z,t)\pi_\theta(1|x,z,t) + G_0(x,z,t)\pi_\theta(0|x,z,t)\right].
\end{aligned}$$

With known weights, one can extend the methods of Athey and Wager (2018) to show that equation (5.3) still holds. Consequently, Theorem 2 continues to hold true.

In reality, $w_t(\cdot)$ is unknown as the distribution F is unknown. We shall assume however that we have access to covariate specific forecasts, $\lambda_x(t)$. We then suggest approximating $w_t(\cdot)$ with a piece-wise constant function $\hat{w}_t(x)$ by partitioning the space $\mathcal{X}$ of the covariates into a finite set of clusters $j = 1, \ldots, J$. The value of $\hat{w}_t(\cdot)$ is constant within each cluster. Ideally, we would like to obtain time varying clusters by clustering on $\lambda_x(t)$, for each given t. This is of course computationally infeasible since t is a continuous variable, so we will instead assume that the heterogeneity in arrival rates is driven by a low dimensional latent variable κ, of dimension d_κ. Then the clusters can be determined by partitioning on a vector of moments of $\lambda_x(t)$, e.g $(\int \lambda_x(t)dt, \int \lambda_x^2(t)dt)$ etc., assuming they are injective with respect to κ (see Bonhomme, Lamadon and Manresa, 2017). Note that the resulting clusters do not change with time.

Denote $\mathcal{X}(j)$ as the domain of cluster j. With the knowledge of the clusters, we can calculate the cluster-specific arrival rate $\lambda_j(t) = \int_{\mathcal{X}(j)} \lambda_x(t)dx$. The value of $\hat{w}_t(\cdot)$ within each cluster j, denoted by $\hat{w}_t(j)$, is then obtained as $\hat{w}_t(j) = \lambda_j(t)/\sum_j \lambda_j(t)$. The empirical counterpart of F_t is now $F_{n,t} = n^{-1}\sum_j \hat{w}_t(j)F_{n,j}$, where $F_{n,j}$ denotes the empirical distribution of observations for each cluster. Using these quantities, the empirical PDE (3.10) is constructed by setting

$$\begin{aligned}\hat{r}_\theta(z,t) &= E_{x\sim F_{n,t}}\left[\pi_\theta(1|x,z,t)\hat{r}(x,1)\right],\\ \hat{G}_\theta(z,t) &= E_{x\sim F_{n,t}}\left[G_1(x,z,t)\pi_\theta(1|x,z,t) + G_0(x,z,t)\pi_\theta(0|x,z,t)\right], \text{ and}\\ \lambda(t) &= \sum_j \lambda_j(t).\end{aligned}$$

The regret rates will now have an additional term due to the approximation error from replacing $w_t(x)$ with the cluster estimate $\hat{w}_t(x)$. Let us denote this rate by $\mathcal{R}(n,J)$, defined as the upper bound

$$\sup_{x\in\mathcal{X}}\left|F_t(x) - \int_{y\le x}\hat{w}_t(y)dF(y)\right| \lesssim \mathcal{R}(n,J).$$

For the clustering scheme described above, we have $\mathcal{R}(n,J) \lesssim J^{-d_\kappa} + \sqrt{J/n}$ (see, e.g, Bonhomme, Lamadon and Manresa, 2017). Different rates are possible under various assumptions and clustering schemes; we shall not document these here but simply use $\mathcal{R}(n,J)$ to state our results. The concentration inequalities (5.3) will now include this as an additional term:[24]

$$E\left[\sup_{z,t\in\mathcal{U},\theta\in\Theta}\|\hat{r}_\theta(z,t) - \bar{r}_\theta(z,t)\|\right] \le C_0\sqrt{\frac{v_1}{n}} + \mathcal{R}(n,J),$$

with related expressions for $\hat{G}_\theta(z,t) - \bar{G}_\theta(z,t)$ and $\hat{\lambda}(t) - \bar{\lambda}(t)$. Subsequently, proceeding as in the proof of Theorem 2 enables us to show that

$$h_{\theta^*}(z_0,t_0) - h_{\hat{\theta}}(z_0,t_0) \le 2C\left(\sqrt{\frac{v}{n}} + \mathcal{R}(n,J)\right),$$

with probability approaching one.

It is straightforward to extend our Actor-Critic algorithm to allow for clusters: before each update we sample the cluster index by drawing the value of j from a multinomial distribution with probabilities $(\hat{w}_t(1),\dots,\hat{w}_t(J))$. The pseudo-code is provided in Appendix B. Also, we have assumed so far that $\lambda_x(t)$ is given. In Appendix F, we present an iterative scheme for estimating the clusters c and cluster specific arrival rates $\hat{\lambda}_c(\cdot)$ jointly from time series data.

6.3. **Deterministic policy rules.** In this section we consider the case where the social planner explicitly wants deterministic rules. Now in examples with substantial functional approximation, the optimal rule is often stochastic (Sutton and Barto, 2018, Chapter 13). Hence, if the goal is

[24] In deriving this expression, we make use of the fact that the concentration bounds in (5.3) hold uniformly over all probability distributions, and therefore uniformly over all $\tilde{F}_t(x) = \int \hat{w}_t(x)dF(x)$. In particular, we can decompose $\hat{r}_\theta(z,t) - \bar{r}_\theta(z,t)$ as the sum of $\hat{r}_\theta(z,t) - \tilde{r}_\theta(z,t)$ and $\tilde{r}_\theta(z,t) - \bar{r}_\theta(z,t)$, where $\tilde{r}_\theta(z,t) := E_{x\sim\tilde{F}_t}[\pi_\theta(1|x,z,t)r(x,1)]$. The first term is the of order $\sqrt{v/n}$ due to the uniform concentration bounds, while the second term $\mathcal{R}(n,J)$ then arises from the difference between $F_t - \tilde{F}_t$ as discussed above.

to obtain a purely deterministic rule, we would often need to explicitly constrain the algorithm to do so.

To illustrate how this can be done, suppose that the social planner is restricted to using generalized linear eligibility scores (Kitagawa and Tetenov, 2018), given by $\mathcal{E} = \{\pi_\theta : \pi_\theta(1|s) = \mathbb{I}(s'\theta > 0)\}$. This functional class is not differentiable. However, we can approximate $\mathcal{E}$ with the soft-max class Π_σ, where $\sigma \in \mathbb{R}^+$ is arbitrarily small. The Actor-Critic algorithm can be directly applied on this approximation class for any fixed σ, but more usefully, we can also treat σ as a 'temperature' parameter, and decrease it slowly in the course of the updates. In doing so we eventually end up with a deterministic policy. In practice, we recommend starting with Algorithm 1 to obtain an initial estimate of θ. After this, the intercept can be extracted and denoted as the initial $1/\sigma$, and the present procedure may be applied.

Our theoretical results require some extensions since the first part of Assumption 1(i) - on the continuity of $\pi_\theta(1|x, z, t)$ - precludes deterministic policies which are usually discontinuous in (z, t) for a given x. However, as noted before, this part of the assumption can be dropped, and we only need that $\bar{G}_\theta(z, t)$ and $\bar{r}_\theta(z, t)$ are uniformly Lipschitz continuous. For instance, the class of linear eligibility scores is discontinuous for any given x, but satisfies Lipschitz continuity of $\bar{G}_\theta(z, t)$ and $\bar{r}_\theta(z, t)$ under some regularity conditions, see Appendix C. A technical difficulty, however, is that $\hat{h}_\theta$ may not exist (as a viscosity solution) since $\hat{G}_\theta(z, t)$ and $\hat{r}_\theta(z, t)$ will still be discontinuous. One way to resolve this is to give up on $\hat{h}_\theta$ entirely and only work with the solution, $\tilde{h}_\theta$, to the approximation scheme (3.13). Theorem 3 then guarantees convergence of $\tilde{h}_\theta$ to h_θ without requiring continuity of $\pi_\theta(1|x, z, t)$.

6.4. **Continuing and online learning.** In this section we discuss how our algorithm may continue to be updated after coming online, as new information is revealed that focuses the forecasts of $\lambda(t)$. Obviously, it is important to keep updating the algorithm if it is intended to be run indefinitely as in Example 1.2. Note that if changes to the forecasts of $\lambda(t)$ are small, it is not computationally too expensive to re-run the program with the new forecasts, starting from the current policy and value function parameters. In general, the optimal policy is continuous in $\lambda(\cdot)$, so we can expect to reach the new optimum within a few episodes. Aggregating over forecasts also adds to the robustness of learned policy against small changes to the forecasts.

We can also let the Actor-Critic algorithm keep updating in the background after coming online. To do so we require the estimates $\hat{r}(s, 1)$ from the observational dataset as we assumed the outcomes are not observable. However, the algorithm now observes the true waiting times between arrivals, and by continuing to update it now implicitly uses the true value of $\lambda(t)$. In this way, the program is able to adjust to changes in arrival rates. The speed of the adjustment will depend on learning rates $(\alpha_{\theta,o}, \alpha_{\nu,o})$ for the online updates. How these should be set will depend on the prior belief over how informative the forecast is, but as a rule of thumb we suggest setting the rates to the same values as used in estimation.[25]

[25]However, it should be noted that the trade-offs involved are somewhat different. With historical data, we can use slower rates than optimal as the drawback is only computational. But with online learning, this would lead to sub-optimal policies. It is therefore important to employ a structured procedure to determine learning rates, such as Population Based Training (see Jaderberg *et al*, 2018).

In settings with infinite horizon, we can also use our algorithm in a completely online manner without historical data, assuming the rewards $Y(1)$ are revealed instantly. Note that contextual bandit algorithms are not applicable here since they do not consider that current actions affect the distribution of the future states (in addition to affecting instantaneous rewards). On the other hand, it can be shown using existing results in Reinforcement Learning that the optimal policy is learnable under our Actor-Critic algorithm, i.e, the program will eventually converge to it assuming the environment is periodic (see e.g, Bhatnagar *et al*, 2009). However the convergence may be extremely slow, leading to substantial welfare losses in the interim. In fact, it would be more efficient to combine both the online and historical learning approaches: at various points, we can periodically pool all the past observations and run an empirical version of the dynamic model (with $\lambda(\cdot)$ estimated from previous observations) to perform additional updates to the policy and value parameters. Doing so speeds up convergence since the model effectively enables us to generate additional observations. The considerations here are essentially the same as that between model-based and model-free reinforcement learning; we refer to Sutton and Barto (2018, Chapter 8) for a discussion.[26]

6.5. **When the policy can affect the environment.** We have assumed so far that the policy does not affect the environment, i.e the distribution and rate of arrival of individuals. However, even if this assumption is not tenable, the results of this paper can be applicable if we have a model of individual response to policy. In particular, if we know how the arrival rates $\lambda(\cdot)$, and the distribution $F(\cdot)$ change with the state variables s, we can incorporate this information into our procedure, and our theoretical results will continue to hold. We illustrate this in Appendix H with a simple example where the distribution of the arrivals depends on the types of people who have been treated before (e.g, men are more likely to apply if the program had historically favored men etc.).

The main difficulty then is to estimate the response to policy. In examples with a finite horizon this is in fact not estimable and one would need to exploit some prior knowledge about the policy response. The situation is however different in infinite horizon setups. Here our algorithm can be used in an online manner and it will eventually learn the optimal policy even in the presence of behavioral response.

7. Empirical application: JTPA

To illustrate our approach, we use the popular dataset on randomized training provided under the JTPA, akin to e.g Kitagawa and Tetenov (2018), or Abadie, Angrist, and Imbens (2002). During 18 months, applicants who contacted job centers after becoming unemployed were randomized to either obtain support or not. Local centers could choose to supply one of the following forms of support: training, job-search assistance, or other support. Again akin to Kitagawa and Tetenov (2018), we consolidate all forms of support. Baseline information about the 20601 applicants was collected as well as their subsequent earnings for 30 months. We follow the sample selection procedure of Kitagawa and Tetenov (2018) and delete entries with missing

[26]Indeed, some of the best performing RL algorithms such as Monte-Carlo-Tree-Search used in AlphaGo (Silver *et al*, 2017) combine both model-based and model-free approaches for a more efficient algorithm.

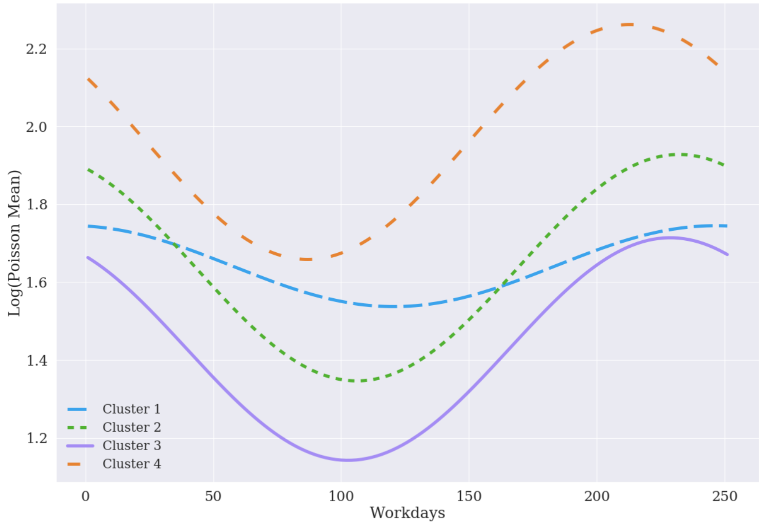


FIGURE 7.1. Clusters-Specific Arrival Rates over Time

earnings or education variables as well as those that are not in the analysis of the adult sample of Abadie, Angrist, and Imbens (2002). This results in 9223 observations.

In this setting, a policy maker is faced with a sequence of individuals who just became unemployed. For each arriving individual, she has to decide whether to offer job training to them or not. The decision is made based on current time, remaining budget, and individual characteristics. For the latter, we follow Kitagawa and Tetenov (2018) and use education, previous earnings, and age. Job training is free to the individual, however, costly to the policy maker who has only limited funds.

The frequency with which people with given characteristics apply is not constant throughout the year. As we use RCT data which contains information regarding when participants arrived, we can estimate Poisson processes that are changing over the course of the year. We first partition the data into clusters using k-median clustering on education, previous earnings, and age. The resulting clusters are briefly described in Appendix G. For each cluster, we estimate the arrival probabilities. While we assume that they are constant across years, we allow for variation within a year. In particular, we specify the following functional form for the cluster-specific Poisson parameter: $\lambda_c(t) = \exp\{\beta_{0,c} + \beta_{1,c} sin(2\pi t) + \beta_{2,c} cos(2\pi t)\}$, where t is normalized so that $t = 1$ corresponds to a year. For each cluster, we obtain the estimates β_c using maximum likelihood (see Appendix F). Figure 7.1 shows the estimated dynamic behavior of each cluster. People from cluster 1, for example, display a less pronounced seasonal pattern regarding their arrival rates than people from cluster 2.

We obtain the reward estimates $\hat{r}(x, 1)$ in two ways: (i) $\hat{r}(x, 1) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$, and (ii) from a doubly robust procedure as in (5.2) that also employs crossfitting. In both cases we use simple OLS to estimate the conditional means. For this reason we shall call case (i) the case of standard OLS rewards. The relevant covariates are education, previous earnings, and age. Estimating the propensity score is not necessary in this context as it was set by the RCT to be $\frac{2}{3}$. Note that the different reward estimates give rise to different heterogeneity patterns, which crucially affect the resulting policy function. Indeed, while the doubly robust procedure consistently estimates the true heterogeneity structure, the standard OLS does not. Consequently, we expect differing parameters in the policy functions and treatment decisions.

In terms of the other parameters, we set the budget such that 1600 people can be treated, which is about a quarter of the expected number of people arriving in a year (given our Poisson rates). This is achieved by normalizing z such that $z_0 = 0.25$ and the cost of treatment to $c = \frac{1}{6400}$. We also use a discount factor of $\beta = -\log(0.9)$, which implies an annualized discount rate of 0.9 (since $t = 1$ corresponds to an year). The episode terminates when all budget is used up.

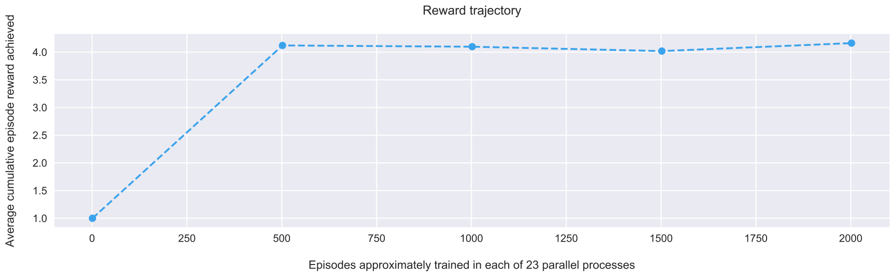
We chose the policy function class such that $\log(\pi_\theta(1|s)/(1-\pi_\theta(1|s)) = \theta_0 + \theta_1^\intercal \mathbf{x} + \theta_2^\intercal \mathbf{x} \cdot z + \theta_3^\intercal \mathbf{x} \cdot \cos(2\pi t)$, where $\mathbf{x} = (1, \text{age}, \text{education}, \text{previous earnings})$.[27] We use $\cos(2\pi t)$ to ensure that the arrival rates are periodic, and to prevent discontinuities at the end of the year. Note that this allows for episodes potentially lasting longer than a year, but constrains the years themselves to be identical.

To run our Actor-Critic algorithm we need to set the learning rates. We tuned these manually starting from the rules of thumb to optimize the performance of the algorithm. Based on pilot runs we found that by setting $\alpha_\theta = 0.3$ and $\alpha_\nu = 0.6$ we could achieve good performance. Moreover, our implementation has 23 reinforcement learning agents training in parallel threads and we have set the batch size to $B = 1024$.

Employing the parallel actor-critic algorithm with clusters (see Appendix G and section 6.2) provides promising results. Figure 7.2 shows that the expected welfare converges as learning occurs through the episodes. Each point in Figure 7.2 is an average over 500 evaluation episodes. We normalize welfare so that choosing a random policy provides a welfare of 1. Eventually, the welfare is around three to four times higher than that under random treatment in the initial episode. For the OLS rewards we initialized the policy and value functions by setting all the coefficients to 0. The evaluated welfare quickly increases and reaches a constant level. On the other hand, the parameters in the policy function do not seem to have entirely converged. We suspect this is because the function $\hat{h}_\theta(z, t)$ is flat around the optimum. We also increased the number of episodes to 20000 but this did not affect the welfare by any significant amount.

In the case of doubly robust rewards, we initialize the policy function parameter with the results obtained from the standard OLS. This is done in order to obtain an initial policy that is sufficiently exploratory. The rewards under doubly robust estimates are much more variable. Thus when initialized with a random policy, the program takes too long a time to explore sufficiently, and as a consequence, the parameters take too long to converge. It is therefore preferable to start with a policy that explores until a few years straightaway (the random policy, by contrast, finishes in only 6 months).

[27] The value basis function used is $\phi(z, t) = \left(z, z\sin(2\pi t), z\cos(2\pi t), z^2\cos(2\pi t), z^2, z^3, z^4\right)^\intercal$.



A: Doubly Robust

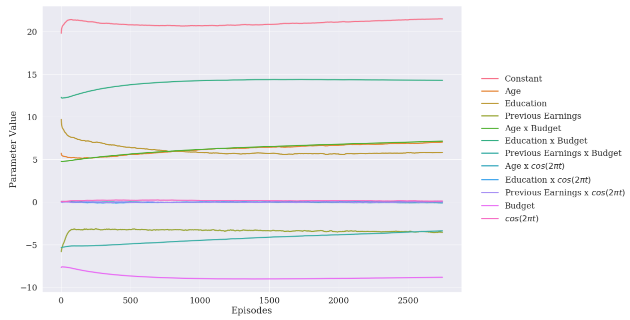
B: Standard OLS

FIGURE 7.2. Converging Episodic Welfare

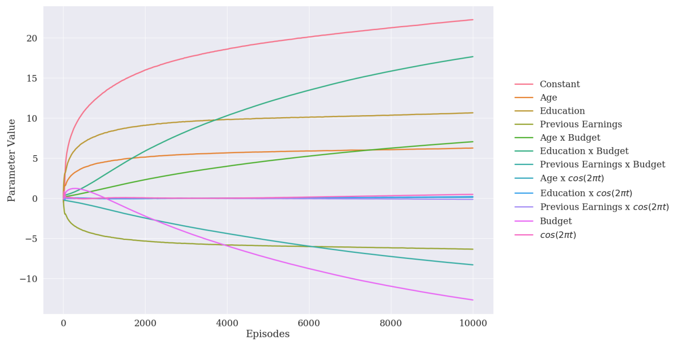
To visualize how the policy function changes with (z, t), we add up all the terms corresponding to a particular covariate, resulting in an expression of the form $\theta_{1(x)} + \theta_{2(x)}z + \theta_{3(x)}\cos(2\pi t)$, where $(\theta_{1(x)}, \theta_{2(x)}, \theta_{3(x)})$ are the coefficients corresponding to covariate x. We then plot these functions as a heatmap for each covariate in Figure 7.4. Note that a low budget indicates that the end of an episode is near, while time reflects differences in seasonal patterns related to the non-constant arrival rates. An episode lasts for approximately three years in the case with standard OLS reward estimates and around 25 years with doubly robust rewards.[28] The heat maps indicate how large the coefficient value is, i.e how strongly this variable influences the decision of treatment. For example, in the case of doubly robust rewards older individuals are more likely to be treated when the remaining budget is still large. Simulations of the resulting policy function allow for further interpretation of the policy function (see Appendix G for details). These simulations show that the policy becomes more selective when the budget decreases. Seasonality, on the other hand, does not seem to matter much in our current example.

In sum, we have shown that rewards resulting from following our estimated policy function are substantially higher than under random treatment. Moreover, both rewards and policy function parameters converge. This illustrates the functionality of our algorithm for given reward estimates. How these estimates are obtained matters: the results for standard OLS rewards are different from those obtained from employing doubly robust estimates (see Appendix G for a

[28]For time running from 0 to 1, the horizontal pattern that the left and right margin are equal to each other with gradual changes towards the middle is artificially due to our choice to have time in the policy function in the form $cos(2\pi t)$.



A: Doubly Robust



B: Standard OLS

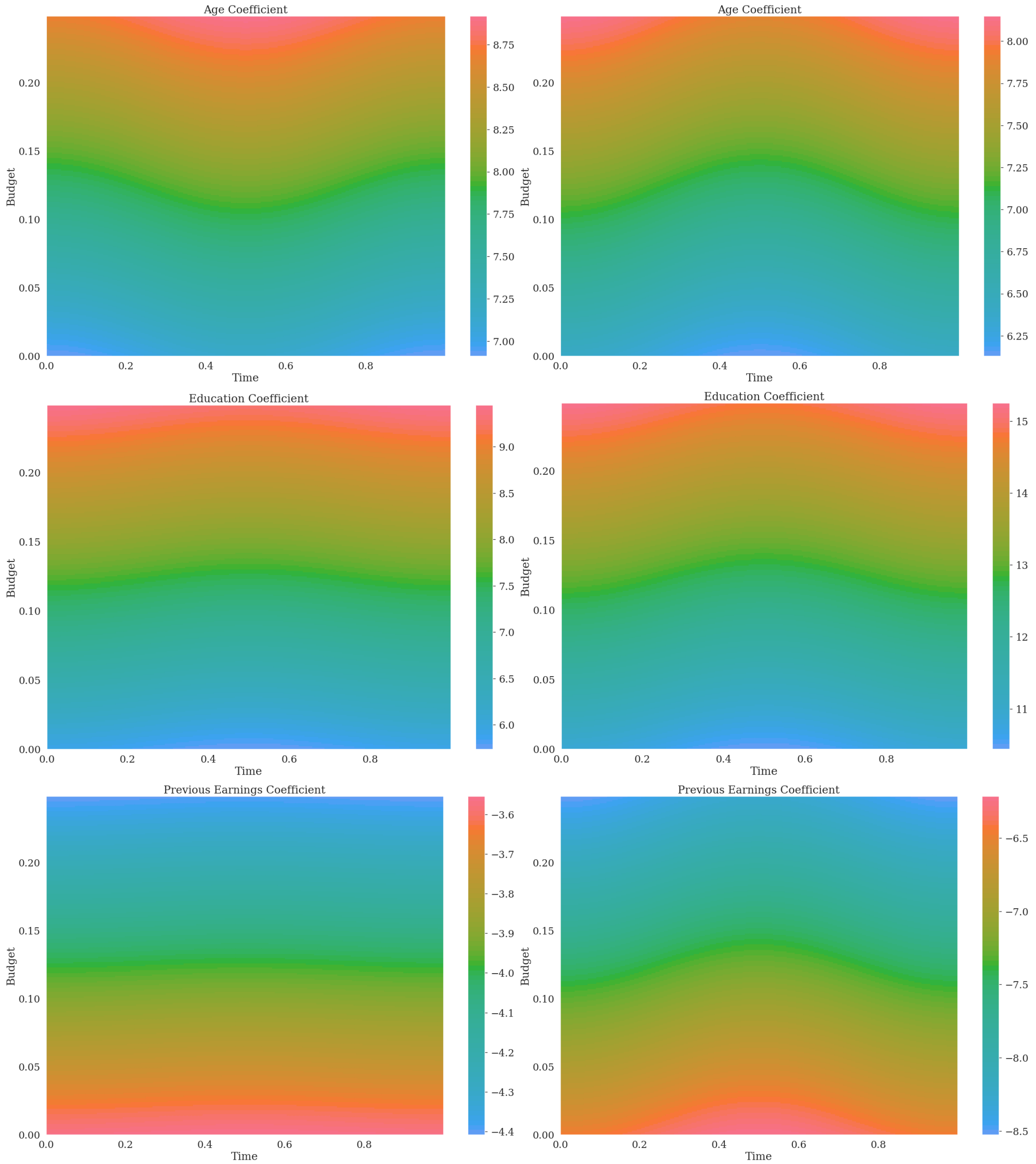
FIGURE 7.3. Convergence of Policy Function Parameters

discussion). Since the latter consistently estimates the true heterogeneity structure, we clearly recommend using doubly robust reward estimation out of the two.

8. CONCLUSION

In this paper we have shown how to estimate optimal dynamic treatment assignment rules using observational data under constraints on the policy space. We proposed an Actor-Critic algorithm to efficiently solve for these rules. Our framework is very general and allows for a broad class of dynamic settings. Separately, our results also point the way to using Reinforcement Learning to solve PDEs characterizing the evolution of value functions. We do so by approximating the PDEs with a dynamic program. We were also able to characterize the numerical error involved in this approximation.

At the same time, the work raises a number of avenues for future research. We have already discussed in previous sections the need to study online learning algorithms, and to allow for strategic behavior on part of individuals. Another drawback of our procedure is that the action space is quite small. For instance, the planner is not allowed to pool individuals and simultaneously allocate treatment given their combined set of covariates. This drawback is also shared by the previous literature on treatment assignment rules. Thus far, the literature has considered static settings where the policy function is determined ex-ante, but is not ex-post optimal conditional on the realized covariates. Our work represents an advance in this regard since we allow the planner to determine the assignment rule as a function of past history. However we



A: Doubly Robust Reward Estimates B: Standard OLS Reward Estimates

FIGURE 7.4. Coefficient Interactions in the Resulting Policy Function

still do not allow the planner to simultaneously decide on a plan of action over a given mass of people. This gives rise to infinite dimensional states and actions, the analysis of which is beyond the scope of the current paper.

References

Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica*, 70(1):91–117.

Achdou, Y., Han, J., Lasry, J.-M., Lions, P.-L., and Moll, B. (2017). Income and wealth distribution in macroeconomics: A continuous-time approach. Technical report, National Bureau of Economic Research.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646.

Athey, S. and Wager, S. (2018). Efficient Policy Learning. *arXiv:1702.02896*.

Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., and Bengio, Y. (2016). An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*.

Barles, G. and Chasseigne, E. (2014). (almost) everything you always wanted to know about deterministic control problems in stratified domains. *arXiv preprint arXiv:1412.7556*.

Barles, G. and Lions, P.-L. (1991). Fully nonlinear neumann type boundary conditions for first-order hamilton–jacobi equations. *Nonlinear Analysis: Theory, Methods & Applications*, 16(2):143–153.

Barles, G. and Souganidis, P. E. (1991). Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic analysis*, 4(3):271–283.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Benitez-Silva, H., Hall, G., Hitsch, G. J., Pauletto, G., and Rust, J. (2000). A comparison of discrete and parametric approximation methods for continuous-state dynamic programming problems. *manuscript, Yale University.*

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. (2009). Natural Actor-Critic Algorithms. *Automatica*, 45(11):2471–2482.

Bhattacharya, D. and Dupas, P. (2012). Inferring Welfare Maximizing Treatment Assignment under Budget Constraints. *Journal of Econometrics*, 167:168–196.

Bonhomme, S., Lamadon, T., and Manresa, E. (2017). Discretizing unobserved heterogeneity. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2019-16).

Bostan, M. and Namah, G. (2007). Time periodic viscosity solutions of hamilton-jacobi equations. In *Applied Analysis And Differential Equations*, pages 21–30. World Scientific.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic Treatment Regimes. *Annual Review of Statistics and Its Application*, 1(1):447–464.

Chamberlain, G. (2011). Bayesian Aspects of Treatment Choice. In Geweke, J., Koop, G., and Van Dijk, H., editors, *The Oxford Handbook of Bayesian Econometrics*, pages 11–39. Oxford University Press.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Crandall, M. G. (1997). Viscosity solutions: a primer. In *Viscosity solutions and applications*, pages 1–43. Springer.

Crandall, M. G., Evans, L. C., and Lions, P.-L. (1984). Some properties of viscosity solutions of hamilton-jacobi equations. *Transactions of the American Mathematical Society*, 282(2):487–502.

Crandall, M. G., Ishii, H., and Lions, P.-L. (1992). User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American mathematical society*, 27(1):1–67.

Crandall, M. G. and Lions, P.-L. (1983). Viscosity solutions of hamilton-jacobi equations. *Transactions of the American mathematical society*, 277(1):1–42.

Crandall, M. G. and Lions, P.-L. (1986). On existence and uniqueness of solutions of hamilton-jacobi equations. *Nonlinear Analysis: Theory, Methods & Applications*, 10(4):353–370.

Crépon, B., Ferracci, M., Jolivet, G., and van den Berg, G. J. (2009). Active labor market policy effects in a dynamic setting. *Journal of the European Economic Association*, 7(2-3):595–605.

Dimakopoulou, M., Athey, S., and Imbens, G. (2017). Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*.

Hirano, K. and Porter, J. R. (2009). Asymptotics for Statistical Treatment Rules. *Econometrica*, 77(5):1683–1701.

Ishii, H. (1985). Hamilton-jacobi equations with discontinuous hamiltonians on arbitrary open sets. *Bull. Fac. Sci. Eng. Chuo Univ*, 28(28):1985.

Kitagawa, T. and Tetenov, A. (2018). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, 86(2):591–616.

Krylov, N. V. (2005). The rate of convergence of finite-difference approximations for bellman equations with lipschitz coefficients. *Applied Mathematics & Optimization*, 52(3):365–399.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225.

Lechner, M. and Wunsch, C. (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics*, 21:111–121.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., and Murphy, S. (2012). A "SMART" Design for Building Individualized Treatment Sequences. *Annual Review of Clinical Psychology*, 8(1):21–48.

Manski, C. F. (2004). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4):1221–1246.

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-Level Control through Deep Reinforcement Learning. *Nature*, 518(7540):529–533.

Murphy, S. A. (2005). An Experimental Design for the Development of Adaptive Treatment Strategies. *Statistics in Medicine*, 24(10):1455–1481.

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471.

Sianesi, B. (2004). An evaluation of the swedish system of active labor market programs in the 1990s. *Review of Economics and statistics*, 86(1):133–155.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354.

Souganidis, P. E. (1985). Existence of viscosity solutions of hamilton-jacobi equations. *Journal of Differential Equations*, 56(3):345–390.

Souganidis, P. E. (2009). Rates of convergence for monotone approximations of viscosity solutions of fully nonlinear uniformly elliptic pde (viscosity solutions of differential equations and related topics).

Stoye, J. (2009). Minimax Regret Treatment Choice with Finite Samples. *Journal of Econometrics*, 151(1):70–81.

Stoye, J. (2012). New Perspectives on Statistical Decisions Under Ambiguity. *Annual Review of Economics*, 4(1):257–282.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063.

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, 166(1):157–165.

Thomas, P. (2014). Bias in natural actor-critic algorithms. In *International Conference on Machine Learning*, pages 441–448.

Tsitsiklis, J. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690.

Vikström, J. (2017). Dynamic treatment assignment and evaluation of active labor market policies. *Labour Economics*, 49:42–54.

Appendix A. Proofs of main results

We recall here the definition of a viscosity solution. Consider a first order differential partial equation of the Dirichlet form

$$F\left(z,t,u(z,t),Du(z,t)\right)=0 \text{ on } \mathcal{U};\quad u=0 \text{ on } \Gamma, \tag{A.1}$$

where x is a vector, Du denotes the derivative with respect to (z,t), $\mathcal{U}$ is the domain of the PDE and $\Gamma\subseteq\partial\mathcal{U}$ is the set on which the boundary conditions are specified. We restrict ourselves to functions $F(\cdot)$ that are *proper*, i.e $F(\cdot)$ is non-decreasing in $u(\cdot,\cdot)$.

In what follows, let $y=(z,t)$ and $y_0=(z_0,t_0)$. Let $\mathcal{Z}$ denote the domain of z. Also, $\mathcal{C}^2(\mathcal{U})$ denotes the space of all twice continuously differentiable functions on $\mathcal{U}$.

Definition 1. *A bounded continuous function u is a viscosity sub-solution to (A.1) if:*

(i) $u\le 0$ on Γ, and

(ii) for each $\phi\in\mathcal{C}^2(\mathcal{U})$, if $u-\phi$ has a local maximum at $y_0\in\mathcal{U}$, then

$$F\left(y_0,u(y_0),D\phi(y_0)\right)\le 0.$$

Similarly, a bounded continuous function u is a viscosity super-solution to (A.1) if:

(i) $u\ge 0$ on Γ, and

(ii) for each $\phi\in\mathcal{C}^2(\mathcal{U})$, if $u-\phi$ has a local minimum at $y_0\in\mathcal{U}$, then

$$F\left(y_0,u(y_0),D\phi(y_0)\right)\ge 0.$$

Finally, u is a viscosity solution to (A.1) if it is both a sub-solution and a super-solution.

We shall also say that u is a viscosity sub-solution to (A.1) *on* $\mathcal{U}$ if only the second condition holds (i.e it need not be the case that $u\le 0$ on Γ). Similarly, u is a viscosity super-solution to (A.1) *on* $\mathcal{U}$ if only condition (ii) holds, without necessarily being the case that $u\ge 0$ on Γ.

The definition of viscosity solutions can also be extended to non-linear boundary conditions following Barles and Lions (1991). Here, we consider a Cauchy problem with a non-linear Neumann boundary condition: (recall that $\mathcal{Z}$ denotes the domain of z)

$$\begin{aligned} F\left(z,t,u(z,t),Du(z,t)\right)&=0 \text{ on } \mathcal{Z}\times(0,\bar{T}];\\ B\left(z,t,u(z,t),Du(z,t)\right)&=0 \text{ on } \partial\mathcal{Z}\times(0,\bar{T}];\\ u(z,t)&=0 \text{ on } \mathcal{Z}\times\{0\}; \end{aligned} \tag{A.2}$$

where $B(\cdot)$ is a non-linear boundary condition. In general, the boundary condition on $\partial\mathcal{Z}\times(t_0,\bar{T}]$ may be over-determined, and the second condition may not hold everywhere. We thus need some weaker notion of the boundary condition as well. This is provided in the following definition, due to Barles and Lions (1991); see also Crandall, Ishii and Lions (1992).

Definition 2. *A bounded continuous function u is a viscosity sub-solution to (A.2) if:*

(i) $u(z,0)\le 0$ for all $z\in\mathcal{Z}$, and

(ii) for each $\phi \in \mathcal{C}^2(\bar{\mathcal{Z}} \times [0,\bar{T}])$, if $u-\phi$ has a local maximum at $y_0 \in \bar{\mathcal{Z}} \times (0,\bar{T}]$, then

$$F\left(y_0, u(y_0), D\phi(y_0)\right) \le 0 \text{ if } y_0 \in \mathcal{Z} \times (0,\bar{T}];$$
$$\min\left\{F\left(y_0, u(y_0), D\phi(y_0)\right), B\left(y_0, u(y_0), D\phi(y_0)\right)\right\} \le 0 \text{ if } y_0 \in \partial\mathcal{Z} \times (0,\bar{T}].$$

Similarly, a bounded continuous function u is a viscosity super-solution to (A.2) if:

(i) $u(z,0) \ge 0$ for all $z \in \mathcal{Z}$, and

(ii) for each $\phi \in \mathcal{C}^2(\bar{\mathcal{Z}} \times [0,\bar{T}])$, if $u-\phi$ has a local minimum at $y_0 \in \bar{\mathcal{Z}} \times (0,\bar{T}]$, then

$$F\left(y_0, u(y_0), D\phi(y_0)\right) \ge 0 \text{ if } y_0 \in \mathcal{Z} \times (0,T];$$
$$\max\left\{F\left(y_0, u(y_0), D\phi(y_0)\right), B\left(y_0, u(y_0), D\phi(y_0)\right)\right\} \ge 0 \text{ if } y_0 \in \partial\mathcal{Z} \times (0,T].$$

Finally, u is a viscosity solution to (A.2) if it is both a sub-solution and a super-solution.

Henceforth, whenever we refer to a viscosity super- or sub-solution, we shall implicitly assume that it is bounded and uniformly continuous. Existence and uniqueness of viscosity solutions for (A.1) and (A.2) can be shown to hold under the regularity conditions given below (see e.g, Crandall, Ishii and Lions, 1992, or Barles and Lions, 1991):

(R1) $F(y,u,p)$ is uniformly continuous in p.

(R2) There exists a modulus of continuity $\omega(\cdot)$ such that

$$|F(y_1,u,p) - F(y_2,u,p)| \le \omega\left(\|y_1-y_2\|\,|1+\|p\||\right).$$

(R3) There exists $\beta > 0$ such that $F(y,u_1,p) - F(y,u_2,p) \ge \beta(u_1-u_2)$ for all $u_1 \ge u_2$.

The regularity conditions on $B(\cdot)$ are very similar, except for one additional condition:

(R4) $B(y,u,p)$ is uniformly continuous in p.

(R5) There exists a modulus of continuity $\omega(\cdot)$ such that

$$|B(y_1,u,p) - B(y_2,u,p)| \le \omega\left(\|y_1-y_2\|\,|1+\|p\||\right).$$

(R6) $B(y,u,p)$ is non-decreasing in u for all (y,p).

(R7) Let $n(y)$ denote the outward normal to Γ at y. There exists $\nu > 0$ such that $B(y,u,p+\lambda n(\cdot)) - B(y,u,p+\mu n(\cdot)) \ge \nu(\lambda-\mu)$ for all $\lambda \ge \mu$.

Finally, we shall also require

(R8) There exists $M \ge 0$ such that $F(y,M,0) \ge 0 \ge F(y,-M,0)$ $\forall (z,t) \in \mathcal{U}$, and $B(y,M,0) \ge 0 \ge B(y,-M,0)$ $\forall (z,t) \in \Gamma$.

A.1. **Proof of Lemma 1.** *Dirichlet boundary condition.* Under the Dirichlet boundary condition, PDE (3.2) can be written as

$$\text{(A.3)} \qquad F_\theta\left(z,t,h_\theta,Dh_\theta\right) = 0 \quad \text{on } \mathcal{U}; \qquad h_\theta(z,t) = 0 \quad \text{on } \Gamma,$$

where $F_\theta(\cdot)$ is defined as

$$F_\theta(z,t,u,p) := \beta u - \left(\lambda(t)\bar{G}_\theta(z,t),1\right)^{\intercal} p - \lambda(t)\bar{r}_\theta(z,t).$$

It is straightforward to verify that the function $F_\theta(\cdot)$ satisfies the regularity conditions (R1)-(R3) and (R8) under Assumption 1. Furthermore, the set Γ as defined in (3.4) satisfies the uniform exterior sphere condition.[29] Then, as long as the above properties are satisfied, the analysis of Crandall (1997, Section 9) shows that a unique viscosity solution exists for (A.3), as long as we are able to exhibit some continuous sub- and super-solutions to (A.3). From the regularity condition (R8), we can see that one such set is given by M and $-M$. Hence a viscosity solution to (A.3) exists.

Periodic boundary condition. We construct the solution to the periodic boundary condition as the long run limit of a Cauchy problem. In particular, let $v_\theta(\cdot)$ denote a solution to the Cauchy problem

$$F_\theta\left(z, t, v_\theta, Dv_\theta\right) = 0 \quad \text{on } \mathbb{R} \times (t_0, \infty);$$
$$v_\theta(z, t) = v_0 \quad \text{on } \mathbb{R} \times \{t_0\},$$

where the function $F_\theta(\cdot)$ is as defined before and v_0 is some arbitrary Lipschitz continuous function e.g $v_0 = 0$. We then claim that if $F_\theta(\cdot)$ is periodic in t, the unique periodic viscosity solution h_θ satisfying (3.5) can be identified as $h_\theta(z,t) = \lim_{m\to\infty} v_\theta(z, mT_p + t)$ for all $t \in [t_0, t_0 + T_p]$. This claim is proved in Bostan and Namah (2007, Proposition 5), but for completeness we restate their arguments here. First, observe that existence of a solution v_θ to the Cauchy problem is assured by our previous arguments. Define $v_\theta^+(z,t) = v_\theta(z, t+T_p)$. By periodicity of $F_\theta(\cdot)$, $v_\theta^+(z,t)$ is a viscosity solution to $F_\theta\left(z, t, v_\theta, Dv_\theta\right) = 0$ on $\mathbb{R} \times (t_0, \infty)$. By Lemma 3 in Appendix D, $|v_\theta| \leq M < \infty$ for some $M < \infty$. Combined with the Comparison Theorem for Cauchy problems (Lemma 5 in Appendix D), we obtain

$$\sup_{(z,w)\in\mathbb{R}\times[t_0,\infty)} |v_\theta^+(z,w) - v_\theta(z,w)| \leq e^{-\beta(w-t)} \sup_{z\in\mathbb{R}} |v_\theta^+(z,t_0) - v_\theta(z,t_0)| \leq 2e^{-\beta(w-t)} M,$$

for any $t < w$. In view of the above equation, setting $w = t + mT_p$, and denoting $h_{m,\theta}(z,t) := v_\theta(z, mT_p + t)$, we have thus shown that

$$\sup_{z,t\in\mathbb{R}\times[t_0,t_0+T_p]} |h_{m+1,\theta}(z,t) - h_{m,\theta}(z,t)| \leq 2e^{-\beta m T_p} M.$$

Thus there exists a limit $h_\theta(z,t)$ to the sequence $h_{m,\theta}(z,t)$. It is clear that this limit is periodic in T_p, as can be seen from the fact

$$|h_{m,\theta}(z, t+T_p) - h_{m,\theta}(z,t)| := |h_{m+1,\theta}(z,t) - h_{m,\theta}(z,t)| \to 0$$

uniformly over all $t \in [t_0, t_0 + T_p]$. Additionally, since $h_{m,\theta}(\cdot)$ is a viscosity solution to $F_\theta\left(z, t, v_\theta, Dv_\theta\right) = 0$ on $\mathbb{R} \times (t_0, \infty)$ for each m, the stability property of viscosity solutions (see Crandall and Lions, 1983) implies that $h_\theta(\cdot)$ is a viscosity solution as well. This completes the existence claim of the periodic solution. That it is also unique follows from the Comparison theorem for periodic boundary condition problems (Theorem 6 in Appendix D).

[29] A set $\mathcal{U}$ is said to satisfy the uniform exterior sphere condition if there exists $r_0 > 0$ such that every point $y \in \partial\mathcal{U}$ is on the boundary of a ball of radius r_0 that otherwise does not intersect $\bar{\mathcal{U}}$.

Neumann and periodic-Neumann boundary conditions. We can rewrite the Neumann boundary condition (3.7) in the form

(A.4)
$$\begin{aligned} F_\theta\left(z,t,h_\theta,Dh_\theta\right) &= 0 \quad \text{on } (\underline{z},\infty)\times[t_0,T);\\ B_\theta\left(z,t,h_\theta,Dh_\theta\right) &= 0 \quad \text{on } \{\underline{z}\}\times(t_0,T);\\ h_\theta(z,t) &= 0 \quad \text{on } [\underline{z},\infty)\times\{T\}, \end{aligned}$$

where

$$B_\theta\left(z,t,u,p\right) := \beta u - \left(\lambda(t)\bar{\sigma}_\theta(z,t),1\right)^\intercal p - \bar{\eta}_\theta(z,t).$$

This can be cast in the form (A.2) after a change of variable $u_\theta(z,\tau) := h_\theta(z,T-\tau)$. Then by the results of Crandall, Ishii and Lions (1992, Theorem 7.12), we can show that a unique solution to (A.4) exists as long as $F_\theta(\cdot)$ and $B_\theta(\cdot)$ satisfy the regularity conditions (R1)-(R8). It is straightforward to verify these under Assumption 1 (note that the outward normal to the plane $\{\underline{z}\}\times[t_0,T)$ is $n=(-1,0)^\intercal$, so (R7) holds as long as $\bar{\sigma}_\theta(z,t)>0$, as assured by Assumption 1(iv)).

For the periodic Neumann boundary condition, we can argue as before by first constructing a solution v_θ to

$$\begin{aligned} F_\theta\left(z,t,v_\theta,Dv_\theta\right) &= 0 \quad \text{on } (\underline{z},\infty)\times[t_0,\infty);\\ B_\theta\left(z,t,v_\theta,Dv_\theta\right) &= 0 \quad \text{on } \{\underline{z}\}\times[t_0,\infty);\\ v_\theta(z,t) &= 0 \quad \text{on } [\underline{z},\infty)\times\{t_0\}, \end{aligned}$$

and then defining $h_\theta(z,t)=\lim_{m\to\infty} v_\theta(z,mT_p+t)$ for $t\in[t_0,t_0+T_p]$.

A.2. **Proof of Theorem 2.** We treat the different boundary conditions separately.

Dirichlet boundary condition. There are two further sub-cases here, depending on whether $T<\infty$ or $T=\infty$. For our proof we choose the case of $T<\infty$. In this setting $\mathcal{U}\equiv(\underline{z},\infty)\times[t_0,T)$, and the boundary condition (3.4) is given by $\Gamma\equiv\{\{\underline{z}\}\times[t_0,T]\}\cup\{(\underline{z},\infty)\times\{T\}\}$, where $\underline{z}\in\mathbb{R}$ (including, potentially, $\underline{z}=-\infty$). We will later sketch how the proof can be modified to deal with the other, arguably simpler, case with $T=\infty$ but finite z, i.e, where $\Gamma\equiv\{\underline{z}\}\times[t_0,\infty)$.

For simplicity, we shall set $t_0=0$. This is without loss of generality. We shall also make a change of variable for t using $\tau(t):=T-t$, which enables us place the boundary condition at $\tau=0$ rather than $t=T$. Define $u_\theta(z,\tau):=e^{\beta\tau}h_\theta(z,T-\tau)$, along with $\hat{u}_\theta(z,\tau):=e^{\beta\tau}\hat{h}_\theta(z,T-\tau)$.[30] In view of (3.2), u_θ satisfies

(A.5)
$$\begin{aligned} \partial_\tau u_\theta + H_\theta(z,\tau,\partial_z u_\theta) &= 0 \quad \text{on } \Upsilon\equiv(\underline{z},\infty)\times(0,T];\\ u_\theta &= 0 \quad \text{on } \mathcal{B}\equiv\{\{\underline{z}\}\times[0,T]\}\cup\{(\underline{z},\infty)\times\{0\}\} \end{aligned}$$

in a viscosity sense, where

$$H_\theta(z,\tau,p) := -e^{\beta\tau}\lambda(\tau)\bar{r}_\theta(z,\tau) - \lambda(\tau)\bar{G}_\theta(z,\tau)p.$$

[30]The multiplication with $e^{\beta\tau}$ allows us to get rid of the term βu_θ, which simplifies the proof. The price we pay however is that we have an additional constant in the rate.

Similarly, from (3.10), $\hat{u}_\theta$ is a viscosity solution to

$$\text{(A.6)}\qquad \begin{aligned} \partial_\tau \hat{u}_\theta + \hat{H}_\theta(z,\tau,\partial_z \hat{u}_\theta) &= 0 \quad \text{on } \Upsilon; \\ \hat{u}_\theta &= 0 \quad \text{on } \mathcal{B}, \end{aligned}$$

where

$$\hat{H}_\theta(z,\tau,p) := -e^{\beta\tau}\lambda(\tau)\hat{r}_\theta(z,\tau) - \lambda(\tau)\hat{G}_\theta(z,\tau)p.$$

We claim that for each $\theta \in \Theta$, $u_\theta(z,\tau) + \tau C\sqrt{v/n}$ is a viscosity super solution to (A.6) on Υ, for some appropriate choice of C. We show this by directly employing the definition of a viscosity super-solution. First, note that $u_\theta(z,\tau) + \tau C\sqrt{v/n}$ is continuous and bounded on $\bar{\Upsilon}$ since so is u_θ (see Lemmas 3 and 4 in Appendix D). Now take any arbitrary point $(z^*,\tau^*) \in \Upsilon$, and let $\phi(z,\tau) \in C^2(\Upsilon)$ be any function such that $u_\theta(z,\tau) + \tau C\sqrt{v/n} - \phi(z,\tau)$ attains a local minimum at (z^*,τ^*). This implies $u_\theta(z,\tau) - \varphi(z,\tau)$ attains a local minimum at (z^*,τ^*), where $\varphi(z,\tau) := -\tau C\sqrt{v/n} + \phi(z,\tau)$. Since $u_\theta(z,\tau)$ is a viscosity solution of (A.5), it follows

$$\partial_\tau \varphi(z^*,\tau^*) + H_\theta\left(z^*,\tau^*,\partial_z\varphi(z^*,\tau^*)\right) \geq 0.$$

The above expression implies

$$\partial_\tau \phi(z^*,\tau^*) - e^{\beta\tau^*}\lambda(\tau^*)\bar{r}_\theta(z^*,\tau^*) - \lambda(\tau^*)\bar{G}_\theta(z^*,\tau^*)\partial_z\phi(z^*,\tau^*) \geq C\sqrt{\frac{v}{n}},$$

and, after a bit more algebra, that

$$\text{(A.7)}\qquad \begin{aligned} &\partial_\tau \phi(z^*,\tau^*) - e^{\beta\tau^*}\lambda(\tau^*)\hat{r}_\theta(z^*,\tau^*) - \lambda(\tau^*)\hat{G}_\theta(z^*,\tau^*)\partial_z\phi(z^*,\tau^*) \\ &\geq C\sqrt{\frac{v}{n}} - e^{\beta\tau^*}\bar{\lambda}\left|\hat{r}_\theta(z^*,\tau^*) - \bar{r}_\theta(z^*,\tau^*)\right| - \bar{\lambda}\left|\hat{G}_\theta(z^*,\tau^*) - \bar{G}_\theta(z^*,\tau^*)\right|\left|\partial_z\phi(z^*,\tau^*)\right| \end{aligned}$$

where $\bar{\lambda} := \sup_\tau \lambda(\tau) < \infty$ by Assumption 1(iii). We shall now show that the right hand side of (A.7) is bounded away from 0, as required by the condition for a viscosity super-solution. To this end, we employ Lemma 4 in Appendix D which assures $u_\theta(\cdot,\tau)$ is Lipschitz continuous in its first argument, with a Lipschitz constant $L_1 < \infty$ independent of z,τ,θ. Consequently, for $u_\theta(z,\tau) - \varphi(z,\tau)$ to attain a local maximum at (z^*,τ^*), it has to be the case that $|\partial_z\varphi(z^*,\tau^*)| \leq L_1$. This in turn implies

$$\text{(A.8)}\qquad |\partial_z\phi(z^*,\tau^*)| \leq L_1.$$

Furthermore, by the results of Athey and Wager (2018), under Assumptions 1-4, there exists a universal constant $C_0 < \infty$ such that

$$\text{(A.9)}\qquad \begin{aligned} \sup_{(z,\tau)\in\bar{\mathcal{U}},\theta\in\Theta} \left|\hat{r}_\theta(z,\tau) - \bar{r}_\theta(z,\tau)\right| &\leq C_0\sqrt{\frac{v_1}{n}}, \text{ and} \\ \sup_{(z,\tau)\in\bar{\mathcal{U}},\theta\in\Theta} \left|\hat{G}_\theta(z,\tau) - \bar{G}_\theta(z,\tau)\right| &\leq C_0\sqrt{\frac{v_2}{n}}, \end{aligned}$$

with probability approaching one (henceforth wpa1). The second inequality in (A.9) can also be derived from Kitagawa and Tetenov (2018, Lemma A.4). In view of (A.7)-(A.9), we can thus set $C > C_0\bar{\lambda}(e^{\beta T} + L_1)$ for which the right hand side of (A.7) is bounded away from 0 wpa1 and we obtain

$$\partial_\tau \phi(z^*, \tau^*) - \lambda(\tau^*)\hat{r}_\theta(z^*, \tau^*) - \lambda(\tau^*)\hat{G}_\theta(z^*, \tau^*)\partial_z \phi(z^*, \tau^*) \geq 0, \quad \text{wpa1}. \tag{A.10}$$

Thus wpa1, $u_\theta(z, \tau) + \tau C\sqrt{v/n}$ is a viscosity super-solution to (A.6) on Υ. This holds true for each $\theta \in \Theta$.

The function $\hat{u}_\theta$ is a viscosity solution, and therefore, a sub-solution to (A.10) on Υ. At the same time, $u_\theta(z, \tau) + \tau C\sqrt{v/n} \geq 0 = \hat{u}_\theta(z, \tau)$ on $\mathcal{B}$ and we have already shown that $u_\theta(z, \tau)+\tau C\sqrt{v/n}$ is a viscosity super solution to (A.6) on Υ. Furthermore, it is straightforward to verify that $\hat{H}_\theta(\cdot)$ satisfies the regularity conditions (H1)-(H3) in Appendix D uniformly in θ, wpa1, in view of Assumption 1 (in particular, the first part of Assumption 1(i) ensures $\hat{G}_\theta(z, t)$ and $\hat{r}_\theta(z, t)$ are uniformly continuous). Consequently, we can apply the Comparison Theorem 5 in Appendix D to conclude

$$\hat{u}_\theta(z, \tau) - u_\theta(z, \tau) \leq \tau C\sqrt{\frac{v}{n}}, \quad \text{wpa1},$$

for all $\theta \in \Theta$. A symmetric argument involving $u_\theta(z, \tau) - \tau C\sqrt{v/n}$ as a sub-solution to (A.10) also implies

$$u_\theta(z, \tau) - \hat{u}_\theta(z, \tau) \leq \tau C\sqrt{\frac{v}{n}}, \quad \text{wpa1},$$

for all $\theta \in \Theta$. Converting the above results back to h_θ and $\hat{h}_\theta$, we obtain

$$\left|\hat{h}_\theta(z, t) - h_\theta(z, t)\right| \leq C(T - t)e^{-\beta(T-t)}\sqrt{\frac{v}{n}}.$$

Since T is finite, this completes the proof of Theorem 2 for the Dirichlet case with a time constraint.

We now briefly sketch how the proof can be modified in the setting with finite z when $T = \infty$. Here $\mathcal{U} \equiv (\underline{z}, z_0] \times [t_0, \infty)$ and $\Gamma \equiv \{\underline{z}\} \times [t_0, \infty)$. Note that assumption 2(ii) implies $\bar{G}_\theta(z, t) < 0$. Then, we make the transformation $u_\theta(z, t) = e^{-\beta t}h_\theta(z, t)$, and write the PDE for $u_\theta(z, t)$ in the form

$$\partial_z u_\theta + H_\theta^{(1)}(t, z, \partial_t u_\theta) = 0 \quad \text{on } \mathcal{U},$$
$$u_\theta = 0 \quad \text{on } \Gamma,$$

where now

$$H_\theta^{(1)}(t, z, p) := e^{-\beta t}\frac{\bar{r}_\theta(z, t)}{\bar{G}_\theta(z, t)} + \frac{p}{\lambda(t)\bar{G}_\theta(z, t)}.$$

The rest of the proof can then proceed as before with straightforward modifications, after reversing the roles of z and t.

Periodic boundary condition. Choose some arbitrary $t^* > T_p$. Denote $u_\theta(z, \tau) = e^{\beta\tau}h_\theta(z, t^* - \tau)$, $\hat{u}_\theta(z, \tau) = e^{\beta\tau}\hat{h}_\theta(z, t^* - \tau)$. Also, set $v_0 := u_\theta(z, 0)$ and $\hat{v}_0 := \hat{u}_\theta(z, 0)$. Now u_θ can be thought of as a viscosity solution to the Cauchy problem

$$\partial_\tau f + H_\theta(z, \tau, \partial_z f) = 0 \quad \text{on } \Upsilon; \tag{A.11}$$
$$f(\cdot, 0) = v_0,$$

where $\Upsilon \equiv \mathbb{R} \times (0, T_p]$ and

$$H_\theta(z, \tau, p) := -e^{\beta\tau}\lambda(\tau)\bar{r}_\theta(z, \tau) - \lambda(\tau)\bar{G}_\theta(z, \tau)p.$$

Similarly, $\hat{u}_\theta(z, \tau)$ is a viscosity solution to

$$\partial_\tau f + \hat{H}_\theta(z, \tau, \partial_z f) = 0 \quad \text{on } \Upsilon; \tag{A.12}$$
$$f(\cdot, 0) = \hat{v}_0,$$

where

$$\hat{H}_\theta(z, \tau, p) := -e^{\beta\tau}\lambda(\tau)\hat{r}_\theta(z, \tau) - \lambda(\tau)\hat{G}_\theta(z, \tau)p.$$

Finally, we shall also define $\tilde{u}_\theta(z, \tau)$ as a viscosity solution to the Cauchy problem

$$\partial_\tau f + \hat{H}_\theta(z, \tau, \partial_z f) = 0 \quad \text{on } \Upsilon; \tag{A.13}$$
$$f(\cdot, 0) = v_0.$$

Note that $\tilde{u}_\theta$ exists and is unique, by the same reasoning as in the proof of Lemma 1. In analogy with the relationship between $u_\theta, \hat{u}_\theta$ and $h_\theta, \hat{h}_\theta$, let us also define

$$\tilde{h}_\theta(z, t) := e^{-\beta t}\tilde{u}_\theta(z, t^* - t).$$

Observe that u_θ and $\tilde{u}_\theta$ share the same boundary condition in (A.11) and (A.13). Furthermore, Lemma 6 in Appendix D assures $u_\theta(\cdot, \tau)$ is Lipschitz continuous in its first argument, with a Lipschitz constant $L_1 < \infty$ independent of z, τ, t, θ. Consequently, we can employ the same arguments as in the Dirichlet setting to show

$$|\tilde{u}_\theta(z, \tau) - u_\theta(z, \tau)| \le C_1\tau\sqrt{\frac{v}{n}}, \quad \text{wpa1},$$

for some constant $C_1 < \infty$ independent of θ, z, τ, t^*. In terms of $\tilde{h}_\theta$ and h_θ, this is equivalent to

$$\left|\tilde{h}_\theta(z, t^* - \tau) - h_\theta(z, t^* - \tau)\right| \le C_1\tau e^{-\beta\tau}\sqrt{\frac{v}{n}}, \quad \text{wpa1}.$$

Setting $\tau = T_p$ in the above expression, and noting that h_θ is T_p-periodic, we obtain

$$\left|\tilde{h}_\theta(z, t^* - T_p) - h_\theta(z, t^*)\right| \le C_1 T_p e^{-\beta T_p}\sqrt{\frac{v}{n}}, \quad \text{wpa1}. \tag{A.14}$$

Now we can also compare $\tilde{u}_\theta$ and $\hat{u}_\theta$ on Υ, using the Comparison Theorem 5 in Appendix D (it is straightforward to note that the regularity conditions are satisfied under Assumption 1). This gives us (henceforth, $(f)_+ := \max\{f, 0\}$)

$$\left(\tilde{u}_\theta(z, T_p) - \hat{u}_\theta(z, T_p)\right)_+ \le \left(\tilde{u}_\theta(z, 0) - \hat{u}_\theta(z, 0)\right)_+, \quad \text{wpa1}.$$

Recall that $\tilde{u}_\theta(z, 0) = v_0 = u_\theta(z, 0)$, by definition. Hence,

$$\left(\tilde{u}_\theta(z, T_p) - \hat{u}_\theta(z, T_p)\right)_+ \le \left(u_\theta(z, 0) - \hat{u}_\theta(z, 0)\right)_+, \quad \text{wpa1}.$$

Rewriting the above in terms of $\tilde{h}_\theta, \hat{h}_\theta$ and h_θ, and noting that $\hat{h}_\theta$ is T_p-periodic, we get

$$e^{\beta T_p}\left(\tilde{h}_\theta(z, t^* - T_p) - \hat{h}_\theta(z, t^*)\right)_+ \le \left(h_\theta(z, t^*) - \hat{h}_\theta(z, t^*)\right)_+, \quad \text{wpa1}. \tag{A.15}$$

In view of (A.14) and (A.15), wpa1,

$$\begin{aligned}\Big(h_\theta(z,t^*) - \hat{h}_\theta(z,t^*)\Big)_+ &\le \Big(\tilde{h}_\theta(z,t^* - T_p) - \hat{h}_\theta(z,t^*)\Big)_+ + C_1 T_p e^{-\beta T_p}\sqrt{\frac{v}{n}} \\ &\le e^{-\beta T_p}\Big(h_\theta(z,t^*) - \hat{h}_\theta(z,t^*)\Big)_+ + C_1 T_p e^{-\beta T_p}\sqrt{\frac{v}{n}}.\end{aligned}$$

Rearranging the above expression gives

$$\Big(h_\theta(z,t^*) - \hat{h}_\theta(z,t^*)\Big)_+ \le C_1 \frac{T_p e^{-\beta T_p}}{1 - e^{-\beta T_p}}\sqrt{\frac{v}{n}}, \quad \text{wpa1.}$$

A symmetric argument - after exchanging the places of $\tilde{u}_\theta$ and $\hat{u}_\theta$ in the lead up to (A.15) - also proves that

$$\Big(\hat{h}_\theta(z,t^*) - h_\theta(z,t^*)\Big)_+ \le C_1 \frac{T_p e^{-\beta T_p}}{1 - e^{-\beta T_p}}\sqrt{\frac{v}{n}}, \quad \text{wpa1.}$$

Since t^* was arbitrary, this concludes the proof of Theorem 2 for the periodic setting.

Neumann boundary condition. We shall recast the Neumann boundary condition problem (3.7) in the form (A.2) by a change of variables through $u_\theta(z,\tau) := e^{\beta\tau} h_\theta(z, T-\tau)$ and $\hat{u}_\theta(z,\tau) := e^{\beta\tau}\hat{h}_\theta(z, T-\tau)$. Note that $u_\theta(z,\tau)$ is a viscosity solution to

$$\begin{aligned} F_\theta(z,\tau,\partial_z u_\theta, \partial_\tau u_\theta) &= 0 \quad \text{on } (\underline{z},\infty)\times(0,T]; \\ B_\theta(z,\tau,\partial_z u_\theta, \partial_\tau u_\theta) &= 0 \quad \text{on } \{\underline{z}\}\times(0,T]; \\ u_\theta(\cdot,0) &= 0, \end{aligned} \tag{A.16}$$

where

$$\begin{aligned} F_\theta(z,\tau,p_1,p_2) &:= -e^{\beta\tau}\lambda(\tau)\bar{r}_\theta(z,\tau) - \lambda(\tau)\bar{G}_\theta(z,\tau)p_1 + p_2, \\ B_\theta(z,\tau,p_1,p_2) &:= -e^{\beta\tau}\bar{\eta}_\theta(z,\tau) - \bar{\sigma}_\theta(z,\tau)p_1 + p_2. \end{aligned}$$

Similarly, $\hat{u}_\theta$ is a viscosity solution to

$$\begin{aligned} \hat{F}_\theta(z,\tau,\partial_z \hat{u}_\theta, \partial_\tau \hat{u}_\theta) &= 0 \quad \text{on } (\underline{z},\infty)\times(0,T]; \\ B_\theta(z,\tau,\partial_z \hat{u}_\theta, \partial_\tau \hat{u}_\theta) &= 0 \quad \text{on } \{\underline{z}\}\times(0,T]; \\ \hat{u}_\theta(\cdot,0) &= 0, \end{aligned} \tag{A.17}$$

where

$$\hat{F}_\theta(z,\tau,p_1,p_2) := -e^{\beta\tau}\lambda(\tau)\hat{r}_\theta(z,\tau) - \lambda(\tau)\hat{G}_\theta(z,\tau)p_1 + p_2.$$

As before, the proof strategy is to show that $u_\theta(z,\tau) + \tau C\sqrt{v/n}$ and $u_\theta(z,\tau) - \tau C\sqrt{v/n}$ are viscosity super- and sub-solutions to (A.17) for some suitable choice of C.

Denote $w_\theta(z,\tau) := u_\theta(z,\tau) + \tau C\sqrt{v/n}$. Clearly, $w_\theta(z,0) = 0 = \hat{u}_\theta(z,0)$. Additionally, using the Lipschitz continuity of u_θ (Lemma 7 in Appendix D), we can recycle the arguments from the Dirichlet setting to show that in a viscosity sense,[31]

$$\hat{F}_\theta(z,\tau,\partial_z w_\theta, \partial_\tau w_\theta) \ge 0 \quad \text{on } (\underline{z},\infty)\times(0,T], \quad \text{wpa1,}$$

[31] It is straightforward to verify that under Assumption 1, the functions $H_\theta(\cdot)$ and $B_\theta(\cdot)$ uniformly satisfy all the regularity conditions (R1)-(R10), as required by the hypothesis of Lemma 7.

for some suitable choice of C. Thus to verify that $w_\theta(z,\tau)$ is a super-solution to (A.17), it remains to show that in a viscosity sense, wpa1,

$$\text{(A.18)} \qquad \max\left\{\hat{F}_\theta(z,\tau,\partial_z w_\theta,\partial_\tau w_\theta), B_\theta(z,\tau,\partial_z w_\theta,\partial_\tau w_\theta)\right\} \geq 0 \text{ on } \{\underline{z}\}\times(0,T].$$

Take an arbitrary point $(\underline{z},\tau^*) \in \{\underline{z}\}\times(0,T]$, and let $\phi(z,\tau) \in C^2([\underline{z},\infty)\times(0,T])$ be any function such that $w_\theta(z,\tau)-\phi(z,\tau)$ attains a local minimum at $(\underline{z},\tau^*)$. We then show below that, wpa1,

$$\text{(A.19)} \qquad \max\left\{\hat{F}_\theta(\underline{z},\tau^*,\partial_z\phi,\partial_\tau\phi), B_\theta(\underline{z},\tau^*,\partial_z\phi,\partial_\tau\phi)\right\} \geq 0,$$

which proves (A.18).

Observe that if $w_\theta(z,\tau)-\phi(z,\tau)$ attains a local minimum at $(\underline{z},\tau^*)$, then $u_\theta(z,\tau)-\varphi(z,\tau)$ attains a local minimum at $(\underline{z},\tau^*)$, where $\varphi(z,\tau) := -\tau C\sqrt{v/n}+\phi(z,\tau)$. Lemma 7 in Appendix D assures u_θ is Lipschitz continuous with constant L_1. Hence, for $(\underline{z},\tau^*)$ to be a local minimum relative to the domain $[\underline{z},\infty)\times[0,T]$, it must be the case [32]

$$\text{(A.20)} \qquad |\partial_\tau\varphi(\underline{z},\tau^*)| \leq L_1, \text{ and } \partial_z\varphi(\underline{z},\tau^*) \leq L_1.$$

Now, by the fact $u_\theta(z,\tau)$ is a viscosity solution of (A.16), we have

$$\max\left\{F_\theta(\underline{z},\tau^*,\partial_z\varphi,\partial_\tau\varphi), B_\theta(\underline{z},\tau^*,\partial_z\varphi,\partial_\tau\varphi)\right\} \geq 0.$$

Suppose $B_\theta(\underline{z},\tau^*,\partial_z\varphi,\partial_\tau\varphi) \geq 0$. Then by $\partial_z\varphi = \partial_z\phi$ and $\partial_\tau\varphi = \partial_\tau\phi - C\sqrt{v/n}$, it is easy to verify $B_\theta(\underline{z},\tau^*,\partial_z\phi,\partial_\tau\phi) \geq C\sqrt{v/n} \geq 0$, which proves (A.19). So let us suppose instead that $B_\theta(\underline{z},\tau^*,\partial_z\varphi,\partial_\tau\varphi) < 0$. We shall use this to obtain a lower bound on $\partial_z\varphi(\underline{z},\tau^*)$. Indeed, $B_\theta(\underline{z},\tau^*,\partial_z\varphi,\partial_\tau\varphi) < 0$ implies

$$\bar{\sigma}_\theta(\underline{z},\tau^*)\partial_z\varphi(\underline{z},\tau^*) > -e^{\beta\tau}\bar{\eta}_\theta(\underline{z},\tau^*) + \partial_\tau\varphi(\underline{z},\tau^*) \geq -C_\eta e^{\beta T} - L_1,$$

where the last inequality follows from Assumption 1(iv) - which ensures $\bar{\eta}_\theta(\underline{z},\tau)$ is bounded above by some constant, say, C_η - and (A.20). But Assumption 1 also assures that $\bar{\sigma}_\theta(\underline{z},\cdot)$ is uniformly bounded away from 0. Hence we conclude $\partial_z\varphi(\underline{z},\tau^*) \geq -L_2$ if $B_\theta(\underline{z},\tau^*,\partial_z\varphi,\partial_\tau\varphi) < 0$, where $L_2 < \infty$ is independent of θ,τ^*. Combined with (A.20), this implies

$$\text{(A.21)} \qquad |\partial_z\varphi(\underline{z},\tau^*)| \leq \max\{L_1,L_2\}, \quad \text{if} \quad B_\theta(\underline{z},\tau^*,\partial_z\varphi,\partial_\tau\varphi) < 0.$$

Now, if $B_\theta(\underline{z},\tau^*,\partial_z\varphi,\partial_\tau\varphi) < 0$ as we supposed, it must be the case $F_\theta(\underline{z},\tau^*,\partial_z\varphi,\partial_\tau\varphi) \geq 0$ to satisfy the requirement for the viscosity boundary condition. Then by similar arguments as in the Dirichlet case, we obtain via (A.21) and (A.9) that [33]

$$\hat{F}_\theta(\underline{z},\tau^*,\partial_z\phi,\partial_\tau\phi) \geq 0, \quad \text{wpa1},$$

as long as $C > C_0(\exp(\beta T) + \bar{\lambda}\max\{L_1,L_2\})$. We have thereby proved (A.19).

[32]Note that it is not necessary $\partial_z\varphi(\underline{z},\tau^*) \geq -L_1$ since $(\underline{z},\tau^*)$ lies on the boundary and we define maxima or minima relative to the domain $[\underline{z},\infty)\times[0,T]$.

[33]In terms of the notation in (A.9), note that here $\bar{\mathcal{U}} \equiv [\underline{z},\infty)\times[0,T]$. Recall also that to get these rates we use Assumption 2 which continuously extends $\bar{r}_\theta$ and $\bar{G}_\theta$ to the boundary.

Returning to the main argument, we have shown by the above that $u_\theta(z,\tau)+\tau C\sqrt{v/n}$ is a super-solution to (A.17), wpa1. At the same time $\hat{u}_\theta(z,t)$ is a sub-solution to (A.17). Furthermore, it is straightforward to verify that $\hat{F}_\theta(\cdot), B_\theta(\cdot)$ satisfy the regularity conditions (R1)-(R10) uniformly in θ, wpa1, in view of Assumptions 1 and (A.9). Hence we can apply the Comparison theorem (7) for the Neumann setting to conclude

$$\hat{u}_\theta(z,\tau)-u_\theta(z,\tau)\le \tau C\sqrt{\frac{v}{n}},\quad \text{wpa1},$$

for all $\theta\in\Theta$. A symmetric argument involving $u_\theta(z,\tau)-\tau C\sqrt{v/n}$ as a sub-solution to (A.17) also implies

$$\hat{u}_\theta(z,\tau)-u_\theta(z,\tau)\le \tau C\sqrt{\frac{v}{n}},\quad \text{wpa1},$$

for all $\theta\in\Theta$. Rewriting the above inequalities in terms of h_θ and $\hat{h}_\theta$, we have thus shown

$$\sup_{(z,t)\in[\underline{z},\infty)\times\{0,T\};\theta\in\Theta}\left|\hat{h}_\theta(z,t)-h_\theta(z,t)\right|\le (T-t)e^{-\beta(T-t)}C\sqrt{\frac{v}{n}}.$$

This concludes our proof of Theorem 2 for the Neumann boundary condition.

Periodic-Neumann boundary condition. This follows from a combination of arguments from the previous cases using Lemma 8 (on Lipschitz continuity of the solution), so we omit the proof.

A.3. **Proof of Theorem 3.** The following proof is based on an argument first sketched by Souganidis (2009) in an unpublished paper.

All the inequalities in this section should be understood to be holding with probability approaching 1 under the distribution F. In what follows, we drop this qualification for ease of notation and hold this to be implicit. We also employ the following notation: For any function f over (z,t), Df denotes its Jacobean. Additionally, $\|\partial_z f\|, \|\partial_z f\|$ and $\|Df\|$ denote the Lipschitz constants for $f(\cdot,t), f(z,\cdot)$ and $f(\cdot,\cdot)$.

We focus here on the Dirichlet boundary condition with $T<\infty$ (but $\underline{z}$ could be $-\infty$). The argument for the other Dirichlet setting, with $T=\infty$ and $\underline{z}>-\infty$, is similar, so we omit it.

We shall represent PDE (3.10) by

$$\begin{aligned} F_\theta(z,t,f,\partial_z f,\partial_t f)&=0, \quad \text{on } \mathcal{U},\\ f&=0, \quad \text{on } \Gamma \end{aligned} \tag{A.22}$$

with f denoting a function, and where

$$F_\theta(z,t,l,p,q):=-\lambda(t)\bar{G}_\theta(z,t)l-p+\beta q-\lambda(t)\bar{r}_\theta(z,t).$$

Additionally, denote our approximation scheme (3.13) by

$$\begin{aligned} S_\theta([f],f,z,t)&=0, \quad \text{on } \mathcal{U},\\ f&=0, \quad \text{on } \Gamma \end{aligned} \tag{A.23}$$

where for any two functions f_1,f_2,

$$S_\theta([f_1],f_2(z,t),z,t,b_n):=b_n\lambda(t)\left(f_2(z,t)-E_{n,\theta}\left[e^{-\beta(t'-t)}f_1(z',t')|z,t\right]\right)-\lambda(t)\hat{r}_\theta(z,t). \tag{A.24}$$

Here $[f]$ refers to the fact that it is a functional argument. Note that h_θ and $\tilde{h}_\theta$ are the functional solutions to (A.22) and (A.23) respectively. We shall also make use of the following two properties for $S_\theta(\cdot)$: First, that $S_\theta(\cdot)$ is monotone in its first argument, i.e

$$S_\theta([f_1], f, z, t, b_n) \geq S_\theta([f_2], f, z, t, b_n) \ \forall \ f_2 \geq f_1. \tag{A.25}$$

Furthermore, for all f and $m \in \mathbb{R}^+$, it holds for all $t \leq T - b_n^{-1/2}$

$$S_\theta([f+m], f+m, z, t, b_n) \geq S_\theta([f], f, z, t) + \chi m, \tag{A.26}$$

where $\chi > \beta/2 > 0$. The first property is trivial to show. As for the second, observe that when t is sufficiently far from the boundary (e.g, $t \leq T - b_n^{-1/2}$),

$$S_\theta([f+m], f+m, z, t, b_n) - S_\theta([f], f, z, t) = m b_n \lambda(t) \left(1 - E_{n,\theta}\left[e^{-\beta(t'-t)}|z, t\right]\right) \geq \chi m.$$

Finally, we recall Lemmas 3, 4 in Appendix D, which show that there exist $K_1, K_2 < \infty$ satisfying

$$\sup_\theta \|h_\theta\| < K_1, \text{ and} \tag{A.27}$$

$$\sup_\theta \|Dh_\theta\| < K_2. \tag{A.28}$$

We provide here an upper bound for

$$m_\theta := \sup_{(z,t)\in\bar{\mathcal{U}}} \left(h_\theta(z,t) - \tilde{h}_\theta(z,t)\right). \tag{A.29}$$

A lower bound for $h_\theta - \tilde{h}_\theta$ can be obtained in an analogous manner. Clearly, we may assume $m_\theta > 0$, as otherwise we are done. Denote (z_θ^*, t_θ^*) as the point at which the supremum is attained in (A.29) (or, if such a point does not exist, where the right hand side of (A.29) is arbitrarily close to m_θ). We shall consider the three (not necessarily mutually exclusive) cases: (i) $|t_\theta^* - T| < 2K\epsilon$, (ii) $|z_\theta^* - \underline{z}| < 2K_2\epsilon$, and (iii) $|z_\theta^* - \underline{z}| \geq 2K_2\epsilon$ and $|t_\theta^* - T| \geq 2K_2\epsilon$. We take ϵ to be any positive number satisfying $\epsilon \geq \sqrt{b_n}$.

We start with Case (i). In view of (A.28), and the fact $h_\theta(z,T) = 0 \ \forall z$, we have

$$|h_\theta(z_\theta^*, t_\theta^*)| < 4K_2^2\epsilon. \tag{A.30}$$

Now, we claim $\tilde{h}_\theta(z,t) \leq L\{(T-t) + b_n^{-1}\}$, for some $L < \infty$ independent of θ, z, t. Let $N[t,T]$ be a random variable denoting the number of arrivals between t and the end point T. Then $N[t,T]$ is first order stochastically dominated by $\bar{N}[t,T] \sim \text{Poisson}(\bar{\lambda} b_n(T-t))$, where $\bar{\lambda} := \sup_t \lambda(t) < \infty$ (note that $\bar{N}[t,T]$ is the number of arrivals between t and T under a Poisson process with parameter $\bar{\lambda} b_n$; the rate of arrivals here is uniformly faster than under the approximation scheme). Hence $E[N[t,T]] \leq E[\bar{N}[t,T]] = \bar{\lambda} b_n(T-t)$. Furthermore, the expected utility gain from any given arrival is atmost $\sup_{\theta,z,t} |\hat{r}_\theta(z,t)|/b_n \leq 2M/b_n$ by Assumption 2(i). Consequently,

$$\tilde{h}_\theta(z,t) \leq \frac{2M}{b_n} + E\left[N[t,T]\frac{2M}{b_n}\right] \leq 2M\bar{\lambda}\left\{(T-t) + b_n^{-1}\right\} := L\left\{(T-t) + b_n^{-1}\right\}.$$

Considering that we are in the case $|t_\theta^* - T| \leq 2K_2\epsilon$, the previous statement implies

$$|\tilde{h}_\theta(z_\theta^*, t_\theta^*)| \leq L\left(2K_2\epsilon + b_n^{-1}\right). \tag{A.31}$$

In view of (A.30) and (A.31), we thus obtain

$$m_\theta \leq (4K_2^2 + 2LK_2)\epsilon + Lb_n^{-1}. \tag{A.32}$$

This completes the treatment of the first case, when $|t_\theta^* - T| < 2K_2\epsilon$. In a similar vein, we can show that (A.32) also holds for Case (ii) using the Lipschitz continuity of h_θ and the fact $\tilde{h}_\theta(z,t) \leq L_2\{|z - \underline{z}| + b_n^{-1}\}$ for some $L_2 < \infty$. The latter follows from $\bar{G}_\theta(z,t) < -\delta$, imposed by Assumption 2(ii), which implies the expected number of arrivals subsequent to state z is bounded above by $\delta^{-1}\{b_n(z - \underline{z}) + C\}$ with $C < \infty$ independent of θ, z, t. The proof of the last statement relies on some martingale results and is deferred to the end of this section.

We now turn to Case (iii), i.e $|z_\theta^* - \underline{z}| \geq 2K_2\epsilon$ and $|t_\theta^* - T| \geq 2K_2\epsilon$. Denote

$$\mathcal{A} \equiv \{(z,t) \in \bar{\mathcal{U}} : \ |z - \underline{z}| \geq 2K_2\epsilon \ \cap \ |t - T| \geq 2K_2\epsilon\}.$$

To obtain the bound on m_θ in this case, we shall employ the sup-convolution of $h_\theta(z,t)$, denoted by $h_\theta^\epsilon(z,t)$:[34]

$$h_\theta^\epsilon(z,t) := \sup_{r,w \in \mathcal{A}} \left\{ h_\theta(r,w) - \frac{1}{\epsilon}\left(|z - r|^2 + |t - w|^2\right) \right\}.$$

In view of (A.28) in Appendix D, and Lemma 10 in Appendix E,

$$\sup_{(z,t)\in\mathcal{A}} |h_\theta(z,t) - h_\theta^\epsilon(z,t)| \leq 4K_2^2\epsilon. \tag{A.33}$$

Also, by Lemma 12 in Appendix E (Assumption 1 ensures all relevant regularity conditions for $F_\theta(\cdot)$ are satisfied), there exists $c < \infty$ independent of θ, z, t such that, in a viscosity sense,

$$F_\theta(z, t, h_\theta^\epsilon, \partial_z h_\theta^\epsilon, \partial_t h_\theta^\epsilon) \leq c\epsilon \quad \text{on } \mathcal{A}. \tag{A.34}$$

Finally, we also note from Lemma 10 in Appendix E that h_θ^ϵ is a semi-convex function with coefficient $1/\epsilon$.[35]

We now compare $S_\theta(\cdot)$ and $F_\theta(\cdot)$ at the function h_θ^ϵ. Consider any $(z,t) \in \mathcal{A}$ at which h_θ^ϵ is differentiable (by semi-convexity, it is differentiable almost everywhere). We can then expand

$$\begin{aligned}
S_\theta([h_\theta^\epsilon], h_\theta^\epsilon, z, t, b_n) &= b_n\lambda(t)h_\theta^\epsilon(z,t)\left(1 - E_{n,\theta}\left[e^{-\beta(t'-t)}|z,t\right]\right) \\
&\quad + b_n\lambda(t)E_{n,\theta}\left[e^{-\beta(t'-t)}\left\{h_\theta^\epsilon(z,t) - h_\theta^\epsilon(z',t')\right\}|z,t\right] - \lambda(t)\hat{r}_\theta(z,t) \\
&:= A_\theta^{(1)}(z,t) + A_\theta^{(2)}(z,t) + A_\theta^{(3)}(z,t).
\end{aligned} \tag{A.35}$$

Using $\|h_\theta^\epsilon\| \leq \|h_\theta\| \leq K_1$ and Assumptions 1-4, straightforward algebra enables us to show

$$A_\theta^{(1)}(z,t) \leq \beta h_\theta^\epsilon(z,t) + \frac{C_1}{b_n}, \tag{A.36}$$

for some C_1 independent of θ, z, t. We now consider $A_\theta^{(2)}(z,t)$. Observe that by semi-convexity of h_θ^ϵ (Lemma 9 in Appendix E),

$$h_\theta^\epsilon(z',t') \geq h_\theta^\epsilon(z,t) + \partial_z h_\theta^\epsilon(z,t)(z'-z) + \partial_t h_\theta^\epsilon(z,t)(t'-t) - \frac{1}{2\epsilon}\left\{|z'-z|^2 + |t'-t|^2\right\}.$$

[34] We discuss sup and inf-convolutions and their properties in Appendix E.

[35] See Appendix E for the definition of semi-convex functions.

Substituting the above into the expression for $A_\theta^{(2)}(z,t)$, and using Assumptions 1, (A.9) and (A.28), some straightforward algebra enables us to show that when $\epsilon \geq \sqrt{b_n}$, [36]

$$A_\theta^{(2)}(z,t) \leq -\lambda(t)\bar{G}_\theta(z,t)\partial_z h_\theta^\epsilon - \partial_t h_\theta^\epsilon + C_2\left(\frac{1}{\epsilon b_n} + \sqrt{\frac{v}{n}}\right), \tag{A.37}$$

where again C_2 is independent of θ, z, t. It remains to bound $A_\theta^{(3)}(z,t)$. For this, we make use of Assumption 1(iii) and (A.9), which ensure there exists C_3 independent of θ, z, t such that

$$A_\theta^{(3)}(z,t) \leq -\lambda(t)\bar{r}_\theta(z,t) + C_3\sqrt{\frac{v}{n}}. \tag{A.38}$$

Combining (A.35)-(A.38), and setting $C = \max(C_1, C_2, C_3)$, we thus find

$$S_\theta([h_\theta^\epsilon], h_\theta^\epsilon, z, t, b_n) \leq F_\theta(z, t, h_\theta^\epsilon, \partial_z h_\theta^\epsilon, \partial_t h_\theta^\epsilon) + C\left\{\frac{1}{b_n}\left(1 + \frac{1}{\epsilon}\right) + \sqrt{\frac{v}{n}}\right\}. \tag{A.39}$$

In view of (A.39) and (A.34),

$$S_\theta([h_\theta^\epsilon], h_\theta^\epsilon, z, t, b_n) \leq c\epsilon + C\left\{\frac{1}{b_n}\left(1 + \frac{1}{\epsilon}\right) + \sqrt{\frac{v}{n}}\right\} \quad \text{a.e.} \tag{A.40}$$

where the qualification almost everywhere (a.e.) refers to the points where Dh_θ^ϵ exists.

Let (here $f^+ := \max(f, 0)$)

$$m_\theta^\epsilon := \sup_{(z,t)\in\mathcal{A}} \left(h_\theta^\epsilon(z,t) - \tilde{h}_\theta(z,t)\right)^+,$$

and denote $(\breve{z}_\theta, \breve{t}_\theta)$ as the point at which the supremum is attained (or where the right hand side of the above expression is arbitrarily close to m_θ^ϵ). Now, by definition,

$$h_\theta^\epsilon \leq \tilde{h}_\theta + m_\theta^\epsilon \text{ on } \mathcal{A}.$$

Then in view of the properties (A.25), (A.26) of $S(\cdot)$,

$$\begin{aligned} \chi m_\theta^\epsilon &= S_\theta\left([\tilde{h}_\theta], \tilde{h}_\theta(\breve{z}_\theta, \breve{t}_\theta), \breve{z}_\theta, \breve{t}_\theta, b_n\right) + \chi m_\theta^\epsilon \\ &\leq S_\theta\left([\tilde{h}_\theta + m_\theta^\epsilon], \tilde{h}_\theta(\breve{z}_\theta, \breve{t}_\theta) + m_\theta^\epsilon, \breve{z}_\theta, \breve{t}_\theta, b_n\right) \\ &\leq S_\theta\left([h_\theta^\epsilon], h_\theta^\epsilon(\breve{z}_\theta, \breve{t}_\theta), \breve{z}_\theta, \breve{t}_\theta, b_n\right). \end{aligned} \tag{A.41}$$

Without loss of generality, we may assume h_θ^ϵ is differentiable at $(\breve{z}_\theta, \breve{t}_\theta)$ as otherwise we can move to a point arbitrarily close, given that h_θ^ϵ is differentiable a.e and Lipschitz continuous; in particular, we note that $S_\theta([f], f(z,t), z, t, b_n)$ is continuous in $(z,t) \in \mathcal{U}$ as long as $f(\cdot)$ is Lipschitz continuous. With this in mind, we can combine (A.41) and (A.40) to obtain

$$m_\theta^\epsilon \leq c_1\epsilon + C\left\{\frac{1}{b_n}\left(1 + \frac{1}{\epsilon}\right) + \sqrt{\frac{v}{n}}\right\}, \tag{A.42}$$

where $c_1 = \chi^{-1}c$ and $C_1 = \chi^{-1}C$ are independent of θ, z, t.

[36] To show this, we use the fact that $E_{n,\theta}[b_n(t'-t)|z,t] = \lambda(t)^{-1} + O(b_n^{-1})$ and $E_{n,\theta}[b_n(z'-z)|z,t] = \hat{G}_\theta(z,t) + O(b_n^{-1})$. Here the $O(b_n^{-1})$ terms arise from the requirement that t', z' have to lie within the boundary, i.e below T and above $\underline{z}$ respectively. When $\epsilon \geq \sqrt{b_n}$, the probability of t', z' crossing the boundary is exponentially small, leading to the $O(b_n^{-1})$ remainder. As for the quadratic terms, observe that $E_{n,\theta}\left[(t'-t)^2|z,t\right] \leq (b_n \inf_t \lambda(t))^{-2}$ and $E_{n,\theta}\left[(z'-z)^2|z,t\right] \leq Cb_n^{-2}$ under the assumption $E_{x\sim F}\left[|G_a(x,z,t)|^2\right] \leq C < \infty$.

Hence, in view of (A.33) and (A.42),

$$m_\theta \leq (4K_2^2 + c_1)\epsilon + C_1 \left\{ \frac{1}{b_n}\left(1 + \frac{1}{\epsilon}\right) + \sqrt{\frac{v}{n}} \right\}. \tag{A.43}$$

This completes the derivation of the upper bound for m_θ under Case (iii), i.e when $|z_\theta^* - \underline{z}| \geq 2K_2\epsilon$ and $|t_\theta^* - T| \geq 2K_2\epsilon$.

Finally, in view of (A.32) and (A.43), setting $\epsilon = b_n^{-1/2}$ gives the desired rate.

Bound on expected number of arrivals after z. It remains to show that the expected number of arrivals subsequent to a state with institutional constraint z is bounded by $\delta^{-1}\{b_n(z - \underline{z}) + C\}$, as was needed for the analysis of Case (ii). Denote by $\{\bar{s}_i \equiv (x_i, z_i, t_i, a_i) : i = 1, 2, \dots\}$ the sequence of state variables following a particular state (z, t), and let

$$M_l := \sum_{i=1}^{l} \left\{ G_{a_i}(x_i, z_i, t_i) - \bar{G}_\theta(z_i, t_i) \right\}.$$

Clearly, M_l is a martingale with respect to the filtration $\mathcal{F}_l := \sigma(\bar{s}_{l-1}, \dots, \bar{s}_1, z, t)$. Let $\mathcal{N}[z]$ be the random variable denoting the number of arrivals until the institutional variable goes below $\underline{z}$ (strictly speaking, $\mathcal{N}[\cdot]$ should also depend on t, but we drop this dependence as our results will be valid for all t). Then $\mathcal{N}[z]$ can be interpreted as the stopping time

$$\mathcal{N}[z] = \inf \left\{ l : \sum_{i=1}^{l} G_{a_i}(x_i, z_i, t_i) \leq b_n(\underline{z} - z) \right\}.$$

Now, the martingale differences of M_l are bounded in expectation by the fact (assumed in the statement of Theorem 3) that $\sup_{a,z,t} E_{x\sim F}[|G_a(x, z, t)|] \leq C$. Hence we can apply the Optional Stopping Theorem, which implies

$$E_{n,\theta}\left[M_{\mathcal{N}[z]}\right] = E_{n,\theta}[M_1] = 0.$$

In other words,

$$E_{n,\theta}\left[\sum_{i=1}^{\mathcal{N}[z]} G_{a_i}(x_i, z_i, t_i) - \sum_{i=1}^{\mathcal{N}[z]} \bar{G}_\theta(z_i, t_i) \right] = 0.$$

By Assumption 2(ii), $-\sum_{i=1}^{\mathcal{N}[z]} \bar{G}_\theta(z_i, t_i) \geq \delta\mathcal{N}[z]$. Furthermore, by definition of $\mathcal{N}[z]$,

$$\begin{aligned} E_{n,\theta}\left[\sum_{i=1}^{\mathcal{N}[z]} G_{a_i}(x_i, z_i, t_i) \right] &= E_{n,\theta}\left[\sum_{i=1}^{\mathcal{N}[z]-1} G_{a_i}(x_i, z_i, t_i) \right] + E_{n,\theta}\left[G_{a_{\mathcal{N}[z]}}(x_{\mathcal{N}[z]}, z_{\mathcal{N}[z]}, t_{\mathcal{N}[z]}) \right] \\ &\geq b_n(\underline{z} - z) - C. \end{aligned}$$

The above implies $\delta E_{n,\theta}[\mathcal{N}[z]] \leq b_n(z - \underline{z}) + C$ or $E_{n,\theta}[\mathcal{N}[z]] \leq \delta^{-1}\{b_n(z - \underline{z}) + C\}$.

Periodic boundary condition. The proof of Theorem 3 for the periodic boundary condition follows by the same reasoning. Indeed, due to periodicity, we can restrict ourselves to the domain $\mathbb{R} \times [t_0, t_0 + T_p]$ and reuse the analysis from Case (iii) above to prove the desired claim (in particular, note that we do not have to worry about seperate cases for the boundary).

Appendix B. Psuedo-codes and additional details for the algorithm

This Section consists of three parts. In the first part, we show how our algorithm can be extended to other boundary conditions beyond the Dirichlet boundary condition setting employed in Section 4. In the second part, we give details about the convergence of the Actor-Critic algorithm in Section 4.2, and provide a proof of Theorem 1 in the main text. In the last part, we provide psuedo-codes and some additional discussion for various extensions to the basic algorithm that were proposed in Section 6.

B.1. **The Actor-Critic algorithm under various boundary conditions.** In Section 4, we described our Actor-Critic algorithm for the Dirichlet boundary condition. Here we look at how it extends to other boundary conditions.

B.1.1. *Neumann boundary condition.* We first start with the Neumann boundary condition. The algorithm is very similar to the Dirichlet case, but we have to be mindful about the fact that the behavior changes at the boundary $z = \underline{z}$. The pseudo-code is described in Algorithm 3, where we employ the example with borrowing constraints for concreteness. Recall that in this setting $\bar{r}_\theta(\underline{z}, t) = 0$ and $\bar{G}_\theta(\underline{z}, t) = \sigma(\underline{z}, t)$. Here $\sigma(\underline{z}, t)$ is the rate of inflow of funds with respect to time, which is assumed to be known. When the flow rate is measured with respect to the expected mass of individuals, this becomes $\sigma(\underline{z}, t)/\lambda(t)$.

Note that in this example the policy parameter θ is not updated when $z = \underline{z}$. This is because the policy function does not exist at this point.[37] However the value function parameters ν are updated since the value function is Lipschitz continuous at $\underline{z}$ (see Section D).

B.1.2. *Periodic/Infinite horizon boundary conditions.* Under the periodic and periodic-Neumann boundary conditions, the policy is intended to be applied indefinitely into the future. In the RL literature, this is known as the continuing case. As mentioned in Section 4, an easy way to extend our algorithms to these settings is to artificially add a time constraint $h_\theta(z, T) = 0$, where T is sufficiently large. In other words, we can approximate the infinite horizon problem with a suitably large finite horizon problem. The numerical error due to this can be bounded using the techniques employed in the proof of Lemma 1 in Appendix A. In particular, suppose that $T = mT_p$ for some integer multiple m. Let $h_\theta(z, t; m)$ denote the value function obtained from restricting the program to $t < mT_p$. We then have for all (z, t),

$$|h_\theta(z, t; m) - h_\theta(z, t)| \leq \sum_{\tilde{m}=m}^{\infty} |h_\theta(z, t; \tilde{m}) - h_\theta(z, t; \tilde{m} + 1)| \leq \sum_{\tilde{m}=m}^{\infty} 2Me^{-\beta \tilde{m} T_p}$$
$$= \frac{2M}{1 - e^{-\beta T_p}} e^{-\beta m T_p}.$$

Thus the numerical error decays exponentially fast with respect to m.

A drawback with the above procedure is that the computational time could become very large for high values of T. A heuristic approach, commonly used in the RL literature (see, e.g, Mnih *et al*, 2015), is to get rid of the $e^{-\beta(t-t_0)}$ term in equations (4.4) and (4.5) and let the

[37] Strictly speaking, the planner is forced to choose $a = 0$ when $z = \underline{z}$. One can extend the definition of the policy function by allowing it to be discontinuous at $\underline{z}$. However, we think this interpretation is not too helpful here, and we prefer to think of continuous policy functions that are restricted to the interior of the domain $\mathcal{U}$.

Algorithm 3: Actor-Critic (Neumann boundary condition)

Initialise policy parameter weights $\theta \leftarrow 0$
Initialise value function weights $\nu \leftarrow 0$

Repeat forever:

Reset budget: $z \leftarrow z_0$
Reset time: $t \leftarrow t_0$
$I \leftarrow 1$

While $t < T$ **:**

If $z > \underline{z}$**:**

$x \sim F_n$ (Draw new covariate at random from data)

$a \sim \pi(a|s;\theta)$ (Draw action, note: $s = (x, z, t)$)

$R \leftarrow \hat{r}(s,a)/b_n$ (with $R = 0$ if $a = 0$)

$z' \leftarrow \max\{z + b_n^{-1} G_a(x,z,t), \underline{z}\}$

Elseif $z = \underline{z}$**:**

$R \leftarrow 0$

$z' \leftarrow \underline{z} + b_n^{-1}\sigma(\underline{z},t)/\lambda(t)$ (Rewrite flow rate wrt mass of people)

$\omega \sim \text{Exponential}(\lambda(t))$

$t' \leftarrow t + \omega/b_n$

$\delta \leftarrow R + \mathbb{I}\{t' < T\} e^{-\beta(t'-t)} \nu^\intercal \phi_{z',t'} - \nu^\intercal \phi_{z,t}$ (Temporal-Difference error)

$\theta \leftarrow \theta + \mathbb{I}\{z > \underline{z}\} \alpha_\theta I \delta \nabla_\theta \ln \pi(a|s;\theta)$ (Update policy parameter)

$\nu \leftarrow \nu + \alpha_\nu I \delta \phi_{z,t}$ (Update value parameter)

$z \leftarrow z'$

$t \leftarrow t'$

$I \leftarrow e^{-\beta(t'-t)} I$

algorithm continue indefinitely until convergence. This simply entails setting $I = 1$ in all our Algorithms 1-3. This approach appears to work very well in practice even though it can be shown theoretically that getting rid of the $e^{-\beta(t-t_0)}$ term in this way can introduce some bias (Thomas, 2014).

Yet a third alternative exploits the equivalence between discounting and random stopping. Mathematically, a Markov Decision Problem (MDP) with some discount factor γ is equivalent to an MDP that is undiscounted but terminates with probability $(1-\gamma)$ after each decision point. Thus random stopping enables us to convert a infinite horizon problem into an episodic one with a random horizon. We exploit this equivalence in Algorithm 4. Convergence of this algorithm follows by the same arguments as in the episodic case. In practice, we recommend parallel updates, since its possible to get draws where the episode takes too long to terminate. Note that our previous procedure, with finite T, essentially amounts to deterministic stopping. Compared to this, random stopping is preferable since there is no bias. Additionally the practitioner does not have to select T.

Algorithm 4: Actor-Critic (Periodic boundary condition with random stopping)

Initialise policy parameter weights $\theta \leftarrow 0$
Initialise value function weights $\nu \leftarrow 0$

Repeat forever:

Reset budget: $z \leftarrow z_0$
Reset time: $t \leftarrow t_0$

While break == FALSE :

$x \sim F_n$ (Draw new covariate at random from data)

$a \sim \pi(a|s;\theta)$ (Draw action, note: $s = (x, z, t)$)

$R \leftarrow \hat{r}(s, a)/b_n$ (with $R = 0$ if $a = 0$)

$\omega \sim \text{Exponential}(\lambda(t))$

$t' \leftarrow t + \omega/b_n$

$\gamma \leftarrow \exp\{-\beta(t' - t)\}$ (Continuation probability)

$\tau \sim \text{Binomial}(1 - \gamma)$ (Stopping variable)

$z' \leftarrow z + G_a(x, z, t)/b_n$

$\delta \leftarrow R + (1 - \tau)\nu^\intercal \phi_{z',t'} - \nu^\intercal \phi_{z,t}$ (Temporal-Difference error)

$\theta \leftarrow \theta + \alpha_\theta \delta \nabla_\theta \ln \pi(a|s;\theta)$ (Update policy parameter)

$\nu \leftarrow \nu + \alpha_\nu \delta \phi_{z,t}$ (Update value parameter)

$z \leftarrow z'$

$t \leftarrow t'$

If $\tau == 1$**:**

break == TRUE

B.2. **Convergence of the Actor-Critic algorithm.** Let $\bar{h}_\theta := \bar{\nu}_\theta^\intercal \phi_{z,t}$, where $\bar{\nu}_\theta$ denotes the fixed point of the value function updates (4.5) for any given value of θ. This is the 'Temporal-Difference fixed point', and is known to exist and also to be unique (Tsitsiklis and van Roy, 1997). We will also make use of the quantities

$$\bar{h}_\theta^+(z,t) \equiv E_\theta[\hat{r}_n(s,a) + \mathbb{I}\{(z',t') \in \mathcal{U}\} e^{-\beta(t'-t)} \bar{h}_\theta(z',t')|z,t]$$

and

$$\mathcal{E}_\theta = E_\theta\left[e^{-\beta(t-t_0)}\left\{\nabla_\theta \bar{h}_\theta^+(z,t) - \nabla_\theta \bar{h}_\theta(z,t)\right\}\right].$$

Define $\mathcal{Z}$ as the set of local minima of $J(\theta) \equiv \bar{h}_\theta(z_0, t_0)$, and $\mathcal{Z}^\epsilon$ an ϵ-expansion of that set. Also, $\theta^{(k)}$ denotes the k-th update of θ. The following theorem is a straightforward consequence of the results of Bhatnagar *et al* (2009).

Theorem 4. *(Bhatnagar et al, 2009) Suppose that Assumptions C hold. Then, given* $\epsilon > 0$*, there exists* δ *such that, if* $\sup_k |\mathcal{E}_{\theta^{(k)}}| < \delta$*, it holds that* $\theta^{(k)} \to \mathcal{Z}^\epsilon$ *with probability 1 as* $k \to \infty$*.*

Intuition for the above theorem can be gleaned from the fact that the expected values of updates for the policy parameter are approximately given by

$$E\left[e^{-\beta(t-t_0)}\delta_n(s,s',a)\nabla_\theta \ln \pi(a|s;\theta)\right] \approx \nabla_\theta J(\theta) + \mathcal{E}_\theta.$$

Thus the term $\mathcal{E}_\theta$ acts as bias in the gradient updates. One can show from the properties of the temporal difference fixed point that if $d_\nu = \infty$, then $\bar{h}_\theta(z,t) = \bar{h}_\theta^+(z,t) = \tilde{h}_\theta(z,t)$, see e.g Tsitsiklis and van Roy (1997). Hence, in this case $\mathcal{E}_\theta = 0$. More generally, it is known that

$$\bar{h}_\theta(z,t) = P_\phi[\bar{h}_\theta^+(z,t)],$$

where P_ϕ is the projection operator onto the vector space of functions spanned by $\{\phi^{(j)} : j = 1,\dots,d_\nu\}$. This implies that $\nabla_\theta \bar{h}_\theta^+(z,t) - \nabla_\theta \bar{h}_\theta(z,t) = (I - P_\phi)[\nabla_\theta \bar{h}_\theta](z,t)$[38]. Now, $\nabla_\theta \bar{h}_\theta$ is uniformly (where the uniformity is with respect to θ) Hölder continuous as long as $\nabla_\theta \pi_\theta(s)$ is also uniformly Hölder continuous in s.[39] Hence for a large class of sieve approximations (e.g Trigonometric series), one can show that $\sup_\theta \left\| (I - P_\phi)[\nabla_\theta \bar{h}_\theta] \right\| \le A(d_\nu)$ where $A(.)$ is some function satisfying $A(x) \to 0$ as $x \to \infty$. This implies $\sup_\theta |\mathcal{E}_\theta| \le A(d_\nu)$. The exact form of $A(.)$ depends on the smoothness of $\nabla_\theta \bar{h}_\theta$, and therefore that of $\nabla_\theta \pi_\theta(s)$, with greater smoothness leading to faster decay of $A(.)$. In view of the above discussion, we have thus shown the following:

Corollary 3. *Suppose that Assumptions C hold and additionally that $\nabla_\theta \pi_\theta(s)$ is uniformly Hölder continuous in s. Then, for each $\epsilon > 0$, there exists M such that if $d_\nu \ge M$, then $\theta^{(k)} \to \mathcal{Z}^\epsilon$ with probability 1 as $k \to \infty$.*

The above was stated as Theorem 1 in the main text.

B.3. **Extensions and Pseudo-codes.** Algorithms 5 and 6 provide the pseudo-codes for the algorithm with non-compliance and clusters respectively. Both algorithms are provided for the Dirichlet boundary condition. The treatment of the other boundary conditions follows analogously to Algorithms 3 and 4. We omit these for brevity.

[38] To verify this, note that we can associate $\bar{h}_\theta, \bar{h}_\theta^+$ with vectors and P_ϕ with a matrix since we assumed discrete values for z, t.

[39] This can be shown easily from the definition of the temporal difference fixed point.

Appendix C. Additional details and extensions for Section 3

This Section consists of four parts. In the first part, we provide various informal derivations of PDE (3.2). In the second part, we consider extensions of our setup to vector-valued institutional variables. The third part discusses Assumption 1(i) in greater detail, and provides some suggestions on how this can be weakened. In the last part, we discuss alternative welfare criteria where the welfare is measured relative to the 'true' or actual values of the arrival rates $\lambda(t)$.

C.1. An intuitive derivation of PDE (3.2).

In this section, we provide three informal derivations of PDE (3.2). The first uses a no-arbitrage argument. The second derives it as the limit of a discrete dynamic programming problem with exponential arrivals. The third derives it as the characterization of the value function when the arrivals are given by an inhomogenous Poisson point process with parameter $\lambda(t)N$, and $N \to \infty$.

C.1.1. *No arbitrage.*

Let us first interpret PDE (3.2) in terms of a no-arbitrage argument. Consider an asset whose value is indexed to the integrated value function $h_\theta(z,t)$, and which pays out the flow (i.e. the flow with respect to time) reward $\lambda(t)\bar{r}_\theta(z,t)$ as dividend. The flow return on this asset, at the natural rate of interest β, is given by $\beta h_\theta(z,t)$. By a no-arbitrage argument, the return on this asset has to equal the dividend, i.e the flow rewards, plus the expected rate of change of value of the asset with respect to time, i.e the (expected) total time derivative $dh_\theta(z,t)/dt$. But note that $\lambda(t)\bar{G}_\theta(z,t)$ is the expected flow change, $\langle dz/dt \rangle$, of the institutional variable with respect to time. Hence,

$$\frac{dh_\theta(z,t)}{dt} = \lambda(t)\bar{G}_\theta(z,t)\partial_z h_\theta(z,t) + \partial_t h_\theta(z,t).$$

Thus the no-arbitrage argument implies $\beta h_\theta(z,t) = \lambda(t)\bar{r}_\theta(z,t) + dh_\theta(z,t)/dt$, or equivalently,

$$\beta h_\theta(z,t) = \lambda(t)\bar{r}_\theta(z,t) + \lambda(t)\bar{G}_\theta(z,t)\partial_z h_\theta(z,t) + \partial_t h_\theta(z,t).$$

The above is just a rearrangement of PDE (3.2).

C.1.2. *As the limit of a sequence of discrete dynamic programming problems with exponential arrivals.*

We now provide a second intuitive argument for PDE (3.2) as the limit of a sequence of discrete dynamic programming problems corresponding to an exponential arrival time process. The PDE is obtained in the limit as the average waiting time tends to 0 (or equivalently the rate of arrival of individuals tends to ∞). This is similar to our 'derivation' of the ODE in Section 2.

Suppose that the time between arrivals is distributed as $\Delta t \sim \mathrm{Exp}(\lambda(t)N)$. Recall that from Section 3 that we had normalized $\lambda(t_0) = 1$. The value of N is proportional to the total number of individuals within any given time interval and provides a way to scale the rate of arrivals. In general, its exact value is not important as we will take it to ∞ shortly. Note that the setting is somewhat different from a Poisson process since there is some history dependence: the value of $\lambda(t)$ is determined by the time of the last arrival.

Define $V_\theta(s)$ as the expected discounted value of all normalized the future rewards $r(s,a)/N$, starting from state s, and when the planner chooses actions according to π_θ. It should be interpreted as the expected value function when the social planner *happens* to be at state s.

We shall let $g_{\lambda(t)}(.)$ denote the probability density function of the exponential distribution with parameter $\lambda(t)$. Now, $V_\theta(.)$ can be obtained as the fixed point of the recursive equation:

$$
\begin{aligned}
V_\theta(s) = & \frac{r(s,1)\pi_\theta(1|s)}{N} \\
& + \int e^{-\beta\frac{\omega}{N}} E_{x'\sim F}\left[V_\theta\left(x', z+\frac{G_1(s)}{N}, t+\frac{\omega}{N}\right)\pi(1|s) + \dots\right. \\
& \left.\dots + V_\theta\left(x', z+\frac{G_0(s)}{N}, t+\frac{\omega}{N}\right)\pi(0|s)\right] g_{\lambda(t)}(\omega)d\omega, \quad \text{for } z>0
\end{aligned} \tag{C.1}
$$

together with

$$V_\theta(s) = 0 \quad \text{for } z = 0.$$

The recursive equation for the integrated value function $h_\theta(z,t) = E_{x\sim F}[V_\theta(s)|z,t]$ is given by

$$
\begin{aligned}
h_\theta(z,t) = & \frac{E_{x\sim F}[r(s,1)\pi_\theta(1|s)|z,t]}{N} \\
& + \int e^{-\beta\frac{\omega}{N}} E_{x\sim F}\left[h_\theta\left(z+\frac{G_1(s)}{N}, t+\frac{\omega}{N}\right)\pi(1|s) + \dots\right. \\
& \left.\dots + h_\theta\left(z+\frac{G_0(s)}{N}, t+\frac{\omega}{N}\right)\pi(0|s)\middle| z,t\right] g_{\lambda(t)}(\omega)d\omega, \quad \text{for } z>0
\end{aligned} \tag{C.2}
$$

together with

$$h_\theta(z,t) = 0 \quad \text{for } z = 0.$$

We now consider the behavior of $h_\theta(z,t)$ in the limit as $N \to \infty$. To this end, first subtract $h_\theta(z,t)\int e^{-\beta\frac{\omega}{N}} g_{\lambda(t)}(\omega)d\omega$ from both sides of equation (C.2) and multiply both sides by N. Next use the definition $\bar{r}_\theta(z,t) := E_{x\sim F}[r(s,1)\pi_\theta(1|s)|z,t]$ and $\bar{G}_\theta(z,t) := E_{x\sim F}\left[G_1(s)\pi_\theta(1|s) + G_0(s)\pi_\theta(0|s)|z,t\right]$. Assuming all the quantities are continuously differentiable to all orders and taking the limit as $N \to \infty$ leads (after some re-arrangement of terms) to PDE (3.2).

C.1.3. *As the limit of the value function under a Poisson point process whose frequency of arrivals goes to infinity.* Finally, we provide a third interpretation of (3.2) as the characterization of the value function when the arrivals are modeled as an inhomogenous Poisson process with parameter $\lambda(t)N$, and $N \to \infty$ (i.e the rate of arrivals is almost continuous). In turns out that in some cases, we can even set $N = 1$. This is possible under an infinite horizon setting where there is no boundary condition on z (otherwise boundary issues could make solution discontinuous).

Let $\mathcal{N}[t_1,t_2]$ denote the counting process for the number of arrivals between time points t_1 and t_2. Note that $\mathcal{N}[t_1,t_2]$ is distributed as a Poisson random variable with parameter $N\int_{t_1}^{t_2}\lambda(t)dt$. As in Section 2, we shall use N to normalize the rewards and budget changes, so that the utility is given by $\bar{r}_\theta(z,t)/N$, and the budget increments by $z'-z = G_a(s)/N$. Also, let us define $\delta = 1/N$. We then have that for points (z,t) that are 'sufficiently' far away from the boundary

$$h_\theta(z,t) = E\left[\frac{1}{N}\sum_{i=1}^{\mathcal{N}[t,t+\delta]}\bar{r}_\theta(Z_{T_i},T_i)\middle| z,t\right] + e^{-\beta\delta}E\left[h_\theta(Z_{t+\delta},t+\delta)|z,t\right],$$

where T_i is a random variable denoting the time of the ith arrival time following time t, and Z_T is a random variable denoting the budget at any arbitrary time T. The above equation comes about since there is a random number, $\mathcal{N}[t,t+\delta]$, of individuals who arrive between t and $t+\delta$, with the covariates of each of these individuals drawn from the distribution F. The equation ignores boundary constraints, which is reasonable if $\delta = 1/N$ is small enough that $Z_{t+\delta}$ lies within the boundary with probability close to 1.

We shall assume that $G_a(s)$ is uniformly bounded by some constant G for all possible a, s. This ensures that $|Z_{T_i} - z| \le GN^{-1}\mathcal{N}[t,t+\delta]$ for all $T_i < t+\delta$. Hence, by Lipschitz continuity of $\bar{r}_\theta(\cdot,\cdot)$ in both its arguments (this is implied by Assumption 1(i)),

$$\begin{aligned} &E\left[\frac{1}{N}\sum_{i=1}^{\mathcal{N}[t,t+\delta]} \bar{r}_\theta(Z_{T_i},T_i)\middle| z,t\right] - \lambda(t)\bar{r}_\theta(z,t)\delta \\ &\le E\left[\frac{\mathcal{N}[t,t+\delta]}{N}\right]\bar{r}_\theta(z,t) - \lambda(t)\bar{r}_\theta(z,t)\delta + E\left[\frac{\mathcal{N}[t,t+\delta]}{N}L_r\left|G\frac{\mathcal{N}[t,t+\delta]}{N}+\delta\right|\right] \\ &= O(\delta^2), \end{aligned} \tag{C.3}$$

where L_r denotes the Lipschitz constant for $\bar{r}_\theta(\cdot,\cdot)$, and we have used the facts

$$E\left[\frac{\mathcal{N}[t,t+\delta]}{N}\right] = \int_t^{t+\delta}\lambda(v)dv = \lambda(t)\delta + O(\delta^2), \text{ and}$$

$$E\left[\frac{\mathcal{N}[t,t+\delta]^2}{N^2}\right] \le \frac{1}{N}\int_t^{t+\delta}\lambda(v)dv + \left(\int_t^{t+\delta}\lambda(v)dv\right)^2 = O(\delta^2).$$

Next, assuming that $h_\theta(\cdot)$ has bounded second derivatives, we obtain

$$\begin{aligned} &E\left[h_\theta(Z_{t+\delta},t+\delta) - h_\theta(z,t)|z,t\right] \\ &= \partial_t h_\theta(z,t)\delta + \partial_z h_\theta(z,t)E[Z_{t+\delta}-z|z,t] + C\left\{\delta^2 + E\left[(Z_{t+\delta}-z)^2|z,t\right]\right\} \\ &= \partial_t h_\theta(z,t)\delta + \partial_z h_\theta(z,t)E[Z_{t+\delta}-z|z,t] + O(\delta^2), \end{aligned}$$

since $E\left[(Z_{t+\delta}-z)^2|z,t\right] \le G^2N^{-2}E\left[\mathcal{N}[t,t+\delta]^2\right] = O(\delta^2)$. Now, we may expand

$$E[Z_{t+\delta}-z|z,t] = E\left[\frac{1}{N}\sum_{i=1}^{\mathcal{N}[t,t+\delta]}\bar{G}_\theta(Z_{T_i},T_i)\middle| z,t\right],$$

and by a similar reasoning to that used for (C.3), it follows

$$E[Z_{t+\delta}-z|z,t] = \lambda(t)\bar{G}_\theta(z,t)\delta + O(\delta^2).$$

In view of the above, we thus obtain

$$(1-e^{-\beta\delta})h_\theta(z,t) = \lambda(t)\bar{r}_\theta(z,t)\delta + \lambda(t)\bar{G}_\theta(z,t)\delta + \partial_t h_\theta(z,t)\delta + O(\delta^2).$$

Dividing the above expression by δ and letting $\delta \to 0$ leads to PDE (3.2).

We emphasize that the above ‘derivation’ is only heuristic as $h_\theta(z,t)$ will not in general be differentiable, let alone twice differentiable. However we can make the argument rigorous through the use of inf- and sup-convolutions as used in the proof of Theorem 3. More precisely, we will be able to employ Taylor expansions as above after approximating $h_\theta(z,t)$ with its inf- and sup-convolutions which are twice differentiable.

C.2. **Vector valued institutional variables.** While we only considered scalar z in the main text, it is straightforward to extend the setup to vector valued z. In this case $G_a(s)$ and $\bar{G}_\theta(z,t)$ will both be vectors and the PDE (3.2) will be of the form

$$\beta h_\theta(z,t) - \lambda(t)\bar{G}_\theta(z,t)^\intercal \partial_z h_\theta(z,t) - \partial_t h_\theta(z,t) - \lambda(t)\bar{r}_\theta(z,t) = 0 \text{ on } \mathcal{U}, \tag{C.4}$$

where $\partial_z \cdot$ is to be interpreted as the partial derivative with respect to a vector valued z.

We need to specify appropriate boundary conditions over the domain $\mathcal{U}$ to close the model. In general different components of z may have different boundary conditions, e.g a Dirichlet boundary condition on the first component and a Neumann one on the other. However, the regret rates and other theoretical properties as given by Theorems 2 & 3 continue to apply and can be derived using the same techniques, even if the analysis becomes tedious due to the multiple boundary conditions.

We describe below some examples with vector valued z:

C.2.1. *Joint budget and capacity constraints.* This example illustrates how the different constraints may be combined. Consider a situation in which the planner has a fixed budget B, but additionally also faces capacity constraints. The institutional variables are $z = (z_1, z_2)$ denoting current budget and the current 'free' capacity respectively. The variable z_1 takes values between $[0, B]$ while $z_2$takes values between $[0, C]$, where C denotes the maximal capacity. A value of $z_2 = 0$ implies the capacity is full. The law of motion for budget is given by $\dot{z}_1 = -\mathbb{I}(a=1)c(s)$, where $c(\cdot)$ denotes the cost of treatment. Similarly, $\dot{z}_2 = \mathbb{I}(a=1) - \lambda(t)^{-1}e(z,t)$, where $e(z,t)$ is the rate (wrt t) at which capacity is freed up. Note that we have normalized the measurement of capacity so that it is filled up at the rate 1 when $a = 1$. Taken together, we thus have

$$G_a(s) = \begin{pmatrix} -\mathbb{I}(a=1)c(s) \\ \mathbb{I}(a=1) - \lambda(t)^{-1}e(z,t) \end{pmatrix}.$$

We can then define the quantities $\bar{r}_\theta(z,t)$ and $\bar{G}_\theta(z,t)$ in the same manner as in the main text. The resulting PDE is given by (C.4).

We need to specify the boundary conditions to complete the model. Here, one of them is determined by the budget constraint, since the program ends when the budget is 0. This gives us a Dirichlet boundary condition

$$h_\theta(0, z_2, t) = 0 \text{ on } \{0\} \times [0, C] \times [t_0, \infty). \tag{C.5}$$

At the same time we also have two Neumann type boundary condition due to possibly discontinuous behaviors when the $z_2 = 0$ or when $z_2 = C$. When the free capacity is 0, the planner is not allowed to treat so that we have $\bar{r}_\theta(z_1, 0, t) = 0$ and $\bar{G}_\theta(z_1, 0, t) = -\lambda(t)^{-1}e(z,t)$. Hence at $z_2 = 0$ we have the boundary condition

$$\beta h_\theta(z,t) - \bar{\sigma}_\theta(z,t)^\intercal \partial_z h_\theta(z,t) - \partial_t h_\theta(z,t) = 0, \quad \text{on } (0, B] \times \{0\} \times [t_0, \infty), \tag{C.6}$$

where $\bar{\sigma}_\theta(z,t) := \left(\bar{G}_{1\theta}(z,t), -\lambda(t)^{-1}e(z,t)\right)^\intercal$, with $\bar{G}_{1\theta}(z,t)$ denoting the first component of $\bar{G}_\theta(z,t)$. At the same time when $z_2 = C$, we must have $e(z,t) = 0$, so that we have another

Neumann boundary condition

(C.7) $$\beta h_\theta(z,t) - \bar{\varsigma}_\theta(z,t)^\intercal \partial_z h_\theta(z,t) - \partial_t h_\theta(z,t) = 0, \quad \text{on } (0,B] \times \{C\} \times [t_0,\infty),$$

where $\bar{\varsigma}_\theta(z,t) := \left(\bar{G}_{1\theta}(z,t), \bar{\pi}_\theta(z,t)\right)^\intercal$ and $\bar{\pi}_\theta(z,t) = E_{x\sim F}[\pi_\theta(x,z,t)]$. Thus equations (C.5) - (C.7) together form the set of boundary conditions for this model.

C.2.2. *Multiple Queues.* Here we consider settings in which the social planner can set up more than one queue. Multiple queues generally enable a more efficient allocation of resources since the planner can use the shorter queue for more time intensive cases. We consider here the case with two queues. The institutional variables are z_1, z_2 denoting the length of the two queues. Suppose that individuals exit the two queues at some known rates $e_1(z,t), e_2(z,t)$ with respect to time. The planner's action a consists of assigning the individuals to one of the queues. We shall denote by $a = 0$ the assignment to queue 1, while the assignment to queue 2 is denoted by $a = 1$. Finally, as with the case of a single queue, we may normalize the measures of z_1, z_2 so that taking action $a = 0$ or 1 adds people to the queues at the rate 1. Then the law of motion for z is given by $\dot{z} = G_a(s)$, where

$$G_a(s) = \begin{pmatrix} \mathbb{I}(a=0) - \lambda(t)^{-1} e_1(z,t) \\ \mathbb{I}(a=1) - \lambda(t)^{-1} e_2(z,t) \end{pmatrix}.$$

An important difference from earlier settings is that we can no longer normalize one of the rewards to 0 (note that here all individuals are eventually treated). The reward for assigning an individual to queue 1 is given by $r(x, z_1, t)$, while that for assigning to queue 2 is given by $r(x, z_2, t)$. The rewards reflect the fact that waiting is costly and the cost is a function of the waiting times i.e the queue length. We then define the quantities

$$\bar{r}_\theta(z,t) := E_{x\sim F}[r(x,z_1,t)\pi_\theta(0|s) + r(x,z_2,t)\pi_\theta(1|s)|z,t], \quad \text{and}$$
$$\bar{G}_\theta(z,t) := E_{x\sim F}\left[G_1(s)\pi_\theta(1|s) + G_0(s)\pi_\theta(0|s)|z,t\right],$$

denoting respectively the expected flow utility and the expected flow change to z. The resulting PDE is of the form (C.4) above. It is natural to setup this problem as a periodic one, either with or without nonlinear Neumann boundary conditions at $z_1 = 0$ and $z_2 = 0$. The latter are useful if the planner would like to allow the policy to behave discontinuously between $z_1 = 0$ and $z_1 > 0$ (or similarly a discontinuity around $z_2 = 0$ and $z_2 > 0$).

C.3. **Additional discussion of Assumption 1.** In this section we provide some primitive regularity conditions under which the soft-max policy class (2.3) satisfies Assumption 1(i). Recall that the soft-max class of policy functions is of the form

$$\pi_\theta(1|s) = \frac{\exp(\theta^\intercal f(s)/\sigma)}{1 + \exp(\theta^\intercal f(s)/\sigma)},$$

where $f(\cdot)$ denotes a vector of basis functions over s. Due to the presence of σ, we can normalize θ to some convenient form. For the present purposes, it will be useful to impose $\|\theta\| = 1$. Let Θ, a subset of the unit sphere, denote the parameter space under consideration for θ.

The following conditions are sufficient to show Assumption 1(i):

Assumption R. *(i) $G_a(s)$ and $r(s,1)$ are uniformly bounded. Furthermore, there exists $C < \infty$ such that $E_{x\sim F}[|\nabla_{(z,t)}G_a(s)|] < C$ and $E_{x\sim F}[|\nabla_{(z,t)}r(s,1)|] < C$ uniformly over all $(z,t) \in \mathcal{U}$.*

(ii) There exists $M < \infty$ independent of (x,z,t) such that $|\nabla_{(z,t)}f(s)| \le M$. This can be relaxed to $E_{x\sim F}[|\nabla_{(z,t)}f(s)|] \le M$ if σ is bounded away from 0.

(iii) Either σ is bounded away from 0, or, there exists $\delta > 0$ such that the probability density function of $\theta^\intercal f(s)$ is bounded in the interval $[-\delta, \delta]$ for each $(z,t) \in \mathcal{U}, \theta \in \Theta$.

Assumption R(i) imposes some regularity conditions on $G_a(s)$ and $r(s,1)$. In our empirical example, these quantities do not even depend on (z,t), so the assumption is trivially satisfied there. Assumption R(ii) ensures that $f(s)$ varies smoothly with (z,t). Assumption R(iii) provides two possibilities. If $1/\sigma$ is compactly supported, it easy to see that the derivatives of $\pi_\theta(\cdot|s)$ with respect to (z,t) are bounded, but this constrains the ability of the policy class to approximate deterministic policies. As an alternative, we can require the distribution of $\theta^\intercal f(s)$ to be bounded around 0 for any given (z,t,θ). It is easy to verify this alternative condition holds as long there exists atleast one continuous covariate, the coefficient of θ corresponding to that covariate is non-zero, and the conditional density of that covariate given the others is bounded away from ∞. The case of discrete covariates with $\sigma \to 0$ presents some difficulties and is discussed in the next sub-section.

Proposition 1. *Suppose that assumptions R(i)-R(iii) hold. Then $\bar{G}_\theta(z,t)$ and $\bar{r}_\theta(z,t)$ are Lipschitz continuous uniformly over θ.*

Proof. Define the soft-max function $\xi(w) = 1/(1+e^{-w/\sigma})$, and let $\xi'(\cdot)$ denote its derivative, which is always positive. Observe that

$$\begin{aligned}\nabla_{(z,t)}\bar{G}_\theta(z,t) &= E_{x\sim F}\left[\nabla_{(z,t)}G_a(s)\pi_\theta(a|s)\right] + E_{x\sim F}\left[G_a(s)\xi'(\theta^\intercal f(s))\theta^\intercal\nabla_{(z,t)}f(s)\right]\\ &\le E_{x\sim F}[|\nabla_{(z,t)}G_a(s)|] + LE_{x\sim F}\left[\xi'(\theta^\intercal f(s))\right],\end{aligned}$$

for some $L < \infty$ independent of (z,t,θ), where the inequality follows from Assumptions R(i)-(ii). It thus remains to show $E_{x\sim F}\left[\xi'(\theta^\intercal f(s))\right] < \infty$. Now $\xi'(w) \le e^{-|w|/\sigma}/\sigma$ for all w, so the previous statement clearly holds when σ is bounded away from 0. For the other possibility in Assumption R(iii), let us pick δ as in the assumption, and expand $E_{x\sim F}\left[\xi'(\theta^\intercal f(s))\right]$ as

$$\begin{aligned}E_{x\sim F}\left[\xi'(\theta^\intercal f(s))\right] &\le E_{x\sim F}\left[\xi'(\theta^\intercal f(s))\mathbb{I}\{|\theta^\intercal f(s)| > \delta\}\right] + E_{x\sim F}\left[\xi'(\theta^\intercal f(s))\mathbb{I}\{|\theta^\intercal f(s)| \le \delta\}\right]\\ &:= A_1 + A_2.\end{aligned}$$

Now without loss of generality, we may assume $\delta \ge \sigma\ln(1/\sigma)$, as otherwise σ is bounded away from 0. Then, by the fact $\xi'(w) \le e^{-|w|/\sigma}/\sigma$, we have $A_1(\delta) \le 1$. Additionally, by Assumption R(iii), the probability density function of $\theta^\intercal f(s)$ is bounded by some constant c, so

$$A_2 \le c\int_{-\delta}^{\delta}\xi'(w)dw \le c[\xi(\delta) - \xi(-\delta)] \le 2c.$$

We thus have $E_{x\sim F}\left[\xi'(\theta^\intercal f(s))\right] \le 1 + 2c < \infty$. This proves Lipschitz continuity of $\bar{G}_\theta(z,t)$. The argument for Lipschitz continuity of $\bar{r}_\theta(z,t)$ is similar. □

C.3.1. *Discrete covariates with arbitrary σ.* For purely discrete covariates with $\sigma \to 0$, it will generically be the case that $\bar{G}_\theta(z,t)$ and $\bar{r}_\theta(z,t)$ are discontinuous, except when the policy is independent of (z,t). Nevertheless, depending on the boundary condition, we can allow for some discontinuities and still end up with a Lipschitz continuous solution. For instance, the results of Ishii (1985) imply a Comparison Theorem (akin to Theorem 5 in Section D) can be derived under the following alternative to Assumption 1(i):

Assumption 1a. *Suppose that the boundary condition is either a periodic one, or of the Cauchy form $h_\theta(z,T) = 0\ \forall\ z$. Then we can replace the second part of Assumption 1(i) with the condition: $\bar{G}_\theta(z,t)$ and $\bar{r}_\theta(z,t)$ are integrable in t on $[t_0,T]$ for any (z,θ), and Lipschitz continuous in z uniformly over (t,θ). A similar condition also holds, with the roles of z,t reversed, if the boundary condition is the form $h_\theta(\underline{z},t) = 0\ \forall\ t$.*

The above condition is also sufficient for proving (uniform) Lipschitz continuity of $h_\theta(z,t)$. To see how, consider the Cauchy condition $h_\theta(z,T) = 0\ \forall\ z$. That $h_\theta(z,t)$ is Lipschitz continuous in z follows by the same reasoning as in Lemma 4, after exploiting the Lipschitz continuity of $\bar{G}_\theta(z,t)$ and $\bar{r}_\theta(z,t)$ with respect to z. As for the Lipschitz continuity of $h_\theta(z,t)$ in the second argument, we can argue as in the second part of Lemma 6; note that this only requires the use of a Comparison Theorem. With these results in hand, we can verify our main Theorems 2 and 3 under the weaker Assumption 1a.

The above results are particularly powerful when applied to the simplified example in Section 2. In this case, the only regularity condition we require for $\bar{\pi}_\theta(z)$ and $\bar{r}_\theta(z)$ is that they have to be integrable and uniformly bounded on $[0,z_0]$, and $\bar{\pi}_\theta(z)$ has to be bounded away from 0.

The general case, when $\bar{G}_\theta(z,t)$ and $\bar{r}_\theta(z,t)$ may be discontinuous in both arguments, is more difficult, but we offer here a few comments. Suppose that there are K distinct covariate groups in the population. Then we can create 2^K strata, each corresponding to regions of (z,t) where the (deterministic) policy function takes the value 1 for exactly one particular subgroup from the K groups. In this way, we can divide the space $\mathcal{U}$ into discrete regions, also called stratified domains, within which $\bar{G}_\theta(z,t)$ and $\bar{r}_\theta(z,t)$ are constant (and therefore uniformly Lipschitz continuous). Discontinuities occur at the boundaries between the strata. Under some regularity conditions, Barles and Chasseigne (2014) demonstrate existence and uniqueness of a solution in this context, and also prove a comparison theorem. It is unknown, however, whether this solution is Lispchitz continuous.

C.4. **Alternative Welfare Criteria.** In the main text we have treated the arrival rates $\lambda(\cdot)$ as forecasts and measured welfare in terms of its 'forecasted' value. Here we consider an alternate criterion where welfare is measure using the realized value or true value of $\lambda(\cdot)$, denoted by $\lambda_0(\cdot)$. Recall that the integrated value function under $\lambda_0(\cdot)$ is denoted by $h_\theta(z,t;\lambda_0)$. Under this welfare criterion the optimal choice of θ is given by

$$\theta_0^* = \arg\max_{\theta\in\Theta} h_\theta(z_0,t_0;\lambda_0).$$

To simplify matters assume that we have access to only a single point forecast or estimate of $\lambda_0(\cdot)$, denoted by $\hat{\lambda}(\cdot)$. The extension to density estimates is straightforward, so we do not consider it here. The criterion function $h_\theta(z_0, t_0; \lambda_0)$ is clearly infeasible. However we can use the historical data and the estimate $\hat{\lambda}(\cdot)$ to obtain the empirical counterpart of $h_\theta(z, t; \lambda_0)$ as $\hat{h}_\theta(z, t; \hat{\lambda})$, where $\hat{h}_\theta(\cdot)$ is the solution to PDE (3.10) from the main text when $\lambda(\cdot)$ is replaced with $\hat{\lambda}(\cdot)$. This suggests the following estimator for θ_0^*:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \hat{h}_\theta(z_0, t_0; \hat{\lambda}).$$

Note that $\hat{\theta}$ is exactly the same as in the main text (cf equation 3.11), excepting that we use $\hat{\lambda}(\cdot)$ in place of $\lambda(\cdot)$. Thus the computation of $\hat{\theta}$ is not affected.

In terms of the statistical properties, the main difference is that we now have to take into account the statistical uncertainty between $\hat{\lambda}(\cdot)$ and $\lambda_0(\cdot)$ while calculating the regret. However estimation of $\lambda_0(\cdot)$ is almost always orthogonal to estimation of treatment effects itself (which are used for estimating $\bar{r}_\theta(z,t)$) since the former is based on time series variation while the latter uses the cross-sectional variation in the data. Indeed, the estimate $\hat{\lambda}(\cdot)$ may even be obtained from a completely different and much bigger dataset: e.g, unemployment rates can be estimated using macro level time series data which is usually much bigger than the RCT data needed to estimate treatment effects. These considerations suggest that the regret can be decomposed into two parts: the first dealing with estimation of the treatment effects, and the other with the estimation of $\lambda_0(\cdot)$. Formally, we can upper bound the regret, $\mathcal{R}_0(\hat{\theta})$, under the present welfare criterion as

$$\begin{aligned}
\mathcal{R}_0(\hat{\theta}) &:= h_{\hat{\theta}}(z_0, t_0; \lambda_0) - h_{\theta_0^*}(z_0, t_0; \lambda_0) \\
&= \left\{ h_{\hat{\theta}}(z_0, t_0; \hat{\lambda}) - h_{\theta_0^*}(z_0, t_0; \hat{\lambda}) \right\} + \left\{ h_{\hat{\theta}}(z_0, t_0; \lambda_0) - h_{\hat{\theta}}(z_0, t_0; \hat{\lambda}) + h_{\theta_0^*}(z_0, t_0; \lambda_0) - h_{\hat{\theta}}(z_0, t_0; \hat{\lambda}) \right\} \\
&\le \left\{ h_{\hat{\theta}}(z_0, t_0; \hat{\lambda}) - h_{\theta_0^*}(z_0, t_0; \hat{\lambda}) \right\} + 2 \sup_{\theta \in \Theta} \left| h_\theta(z_0, t_0; \lambda_0) - h_\theta(z_0, t_0; \hat{\lambda}) \right| \\
&:= \mathcal{R}_0^{(I)} + \mathcal{R}_0^{(II)}.
\end{aligned}$$

Note that $\mathcal{R}_0^{(I)}$ and $\mathcal{R}_0^{(II)}$ are stochastically independent if the estimation of treatment effects and $\hat{\lambda}(\cdot)$ are orthogonal to each other. The first term $\mathcal{R}_0^{(I)}$ can be analyzed using the techniques developed so far. Indeed,[40]

$$\mathcal{R}_0^{(I)} \le 2 \sup_{\theta \in \Theta} \left| \hat{h}_\theta(z_0, t_0; \hat{\lambda}) - h_\theta(z_0, t_0; \hat{\lambda}) \right| \le 2C \sqrt{\frac{v}{n}} \quad \text{wpa1}.$$

As for the second term, we can analyze it using the same PDE techniques as that used in the proof of Theorem 2. This gives us

$$\mathcal{R}_0^{(II)} \le C_1 \sup_{t \in [t_0, \infty)} \left| \lambda_0(t) - \hat{\lambda}(t) \right|,$$

where the constant C_1 depends only on (1) the upper bounds $\sup_{\theta \in \Theta; (z,t) \in \bar{\mathcal{U}}} |\bar{G}_\theta(z,t)|$ & $\sup_{\theta \in \Theta; (z,t) \in \bar{\mathcal{U}}} |\bar{r}_\theta(z,t)|$ and (2) the uniform Lipschitz constants for $\bar{G}_\theta(z,t)$ & $\bar{r}_\theta(z,t)$.[41] We

[40] We shall require $\hat{\lambda}(\cdot)$ to be uniformly upper bounded and bounded away from 0. This is clearly satisfied wpa1 if $\hat{\lambda}(\cdot) - \lambda_0(\cdot) = o_p(1)$ and $\lambda_0(\cdot)$ is upper bounded and bounded away from 0.

[41] Assumption 1(i) assures that all these quantities are indeed finite.

particularly emphasize that $\mathcal{R}_0^{(II)}$ is independent of the complexity v of the policy space. It may even be independent of n if the estimate $\hat{\lambda}(\cdot)$ is constructed using a different dataset.

Combining the above, we have thus shown

$$\mathcal{R}_0(\hat{\theta}) \leq 2C\sqrt{\frac{v}{n}} + C_1 \sup_{t\in[t_0,\infty)} \left|\lambda_0(t) - \hat{\lambda}(t)\right|.$$

Thus the regret rates are exactly the same as that derived in the main text, except for an additional term dealing with estimation of $\lambda_0(\cdot)$. However since this additional term is independent of the complexity of the policy space, the alternative welfare criterion offers no additional implication for choosing the policy class. Thus both welfare criteria lead practically to the same results.

Appendix D. Properties of viscosity solutions

In this Section, we collect various properties of viscosity solutions used in the proof of Theorems 2 and 3. A key result is the Comparison theorem that enables one to prove inequalities between viscosity super- and sub-solutions. We break down this section into separate cases for each of the boundary conditions:

D.1. **Dirichlet boundary condition.** Some of the results in this section apply only to Hamiltoninan-like PDEs with a Dirichlet boundary condition

$$\partial_t u + H\left(z,t,u(z,t),\partial_z u(z,t)\right) = 0 \text{ on } \mathcal{U}; \quad u = 0 \text{ on } \Gamma, \tag{D.1}$$

where $\mathcal{U}$ and Γ are of the form given in the main text. The properties we imposed for $F(\cdot)$ in Appendix A are now transferred to $H(\cdot)$. As before, we let $y := (z,t)$.

(H1) $H(y,u,p)$ is continuous in all its arguments.

(H2) There exists a modulus of continuity $\omega(\cdot)$ such that

$$|H(y_1,u,p_1) - H(y_2,u,p_2)| \leq \omega\left(\|y_1-y_2\| + \|p_1-p_2\|\right), \quad \text{and}$$
$$|H(y_1,u,p) - H(y_2,u,p)| \leq \omega\left(\|y_1-y_2\|\,|1+\|p\||\right).$$

(H3) There exists $\beta > 0$ such that $H(y,u_1,p) - H(y,u_2,p) \geq \beta(u_1-u_2)$ for all $u_1 \geq u_2$.

The following Comparison Theorem states that if a function v is a viscosity super-solution and u a sub-solution satisfying $v \geq u$ on the boundary, then it must be the case that $v \geq u$ everywhere on the domain of the PDE. The version of the theorem that we present here combines Crandall, Ishii & Lions (1992, Theorem 3.3), and Crandall & Lions (1986, Theorem 1). We present the theorem in both the usual and Hamiltonian forms. Clearly, the second set of results is implied by the first, but we present both here to provide an easy reference for the applications of the theorem. (Recall the notation $(f)_+ := \max\{f,0\}$).

Theorem 5. ***(Comparison Theorem - Dirichlet form)*** *(i) Suppose that the function $F(\cdot)$ satisfies conditions (R1)-(R3). Let u,v be respectively, a viscosity sub- and super-solution to*

$$F\left(z,t,f(z,t),\partial_z f(z,t),\partial_t f(z,t)\right) = 0 \text{ on } \mathcal{U},$$

where $\mathcal{U}$ is an open set. Then

$$\sup_{\bar{\mathcal{U}}}(u-v)_+ \leq \sup_{\partial\mathcal{U}}(u-v)_+. \tag{D.2}$$

(ii) Suppose that the function $H(\cdot)$ satisfies conditions (H1)-(H3). Let u,v be respectively, a viscosity sub- and super-solution to

$$\partial_t f + H\left(z,t,f(z,t),\partial_z f(z,t)\right) = 0 \text{ on } \mathcal{U},$$

where $\mathcal{U}$ is an open set. Then (D.2) holds. If, alternatively, $\mathcal{U}$ is the of the form $\mathcal{Z}\times(0,T]$, where $\mathcal{Z}$ is any open set, we can replace $\partial\mathcal{U}$ in the statement with $\Gamma \equiv \{\partial\mathcal{Z}\times[0,T]\}\cup\{\mathcal{Z}\times\{0\}\}$.

Note that the above theorem can be applied without regard to the actual boundary condition.

The following useful lemma is taken from Crandall and Lions (1986).

Lemma 2. ***(Crandall and Lions, 1986, Lemma 2)*** *Suppose that the functions $H_1(\cdot)$ and $H_2(\cdot)$ satisfy conditions (H1)-(H3). Suppose further that u, v are respectively a viscosity sub- and super-solution of $\partial_t f + H_1(z, t, f, \partial_z f) = 0$ and $\partial_t f + H_2(z, t, f, \partial_z f) = 0$ on $\Omega \times (0, T]$, where Ω is an open set. Denote $w(z_1, z_2, t) := u(z_1, t) - v(z_2, t)$. Then $w(z_1, z_2, t)$ satisfies*

$$\partial_t w + H_1(z_1, t, u(z_1, t), \partial_{z_1} w) - H_2(z_2, t, v(z_2, t), \partial_{z_2} w) \leq 0 \text{ on } \Omega \times \Omega \times (0, T]$$

in a viscosity sense.

Lemma 3. *Suppose that Assumptions 1-4 hold for the Dirichlet boundary condition (3.3). Then there exists $L_0 < \infty$ independent of θ, z, t such that $|h_\theta(z, t)| \leq L_0$. In addition, for the setting with $T < \infty$, it holds $|h_\theta(z, t)| \leq K|T - t|$ for some $K < \infty$. In a similar vein, for the setting with $\underline{z} > -\infty$ it holds $|h_\theta(z, t)| \leq K_1|z - \underline{z}|$ for some $K_1 < \infty$.*

Proof. Consider first the Dirichlet problem with $T < \infty$. As in the proof of Theorem 2, we make a change of variable and define $u_\theta(z, \tau) := e^{\beta\tau} h_\theta(z, T - \tau)$. This enable us to recast PDE (3.2) in the form (A.5), as used in the proof of Theorem 2. We now claim that $\phi(z, \tau) := K\tau$ is a super-solution to (3.2) on $\mathcal{U}$, for some appropriate choice of K. Indeed, plugging this function into the PDE, we get

$$\partial_\tau \phi + H_\theta(z, \tau, \partial_z \phi) = K - \lambda(\tau)\bar{r}_\theta(z, \tau).$$

The right hand side is greater than 0 as long as we choose $K \geq \sup_{z,\tau} |\lambda(\tau)\bar{r}_\theta(z, \tau)|$ (note that $|\lambda(\tau)\bar{r}_\theta(z, \tau)|$ is uniformly bounded by virtue of Assumption 2(i)). This proves $\phi(z, \tau) := K\tau$ is a super-solution to (A.5) on $\mathcal{U}$. At the same time it is clear that $\phi \geq 0 \geq u_\theta$ on Γ. Hence, by the Comparison Theorem 5, it follows $u_\theta \leq \phi$ on $\bar{\mathcal{U}}$ (it is straightforward to verify the conditions for the Comparison Theorem 5 under Assumptions 1). Note that this also implies $u_\theta \leq KT$ everywhere. Since $h_\theta(z, t) = e^{-\beta(T-t)} u_\theta(z, T - t)$, this completes the proof for the non-stationary case with finite T.

A similar argument switching the roles of z, τ (as we also did in the proof of Theorem 2) proves that $|h_\theta(z, t)| \leq K_1|z - \underline{z}|$. □

Lemma 4. *Suppose that Assumptions 1-4 hold for the Dirichlet boundary condition (3.3). Then there exists $L_1 < \infty$ independent of θ, z, t such that $h_\theta(z, t)$ is locally Lipschitz continuous in both arguments with Lipschitz constant L_1.*[42]

Proof. We split the proof into three cases:

Case (i), wherein $\underline{z} = -\infty$: In this case, the boundary condition (3.3) is equivalent to a Cauchy problem after a change of variable to τ as in the proof of Theorem 2. In particular, the initial value is provided at $\tau = 0$. For the Cauchy problem, we can apply the results of Souganidis (1985, Proposition 1.5) to show that the h_θ is locally Lipschitz continuous.

[42]We say a function f is locally Lipschitz continuous if $|f(z_1) - f(z_2)| \leq L|z_1 - z_2|$ for all $|z_1 - z_2| < \delta$, where $\delta > 0$. Clearly a locally Lipschitz function is also globally Lipschitz if the domain of z is a compact set.

Case (ii), wherein $T = \infty$: In this case too, we can recast (3.3) as a Cauchy problem, with an initial value provided at $z = 0$. Hence we can again apply Souganidis (1985, Proposition 1.5) to prove the claim.

Case (iii), wherein $\underline{z} > -\infty$ and $T < \infty$: We prove here that $h_\theta(\cdot, t)$ is locally Lipschitz continuous in its first argument. That it is also Lipschitz continuous in its second argument follows by a similar reasoning after switching the roles of z and t. As in the proof of Theorem 2, we make a change of variable and define $u_\theta(z, \tau) := e^{\beta\tau} h_\theta(z, T - \tau)$. This enable us to recast PDE (3.2) in the form (A.5), where

$$H_\theta(z, \tau, p) := -e^{\beta\tau}\lambda(\tau)\bar{r}_\theta(z, \tau) - \lambda(\tau)\bar{G}_\theta(z, \tau)p,$$

as specified in that proof. Denote $\delta_\theta(z_1, z_2, \tau) := u_\theta(z_1, \tau) - u_\theta(z_2, \tau)$. Also let $\Upsilon \equiv (\underline{z}, \infty) \times (\underline{z}, \infty) \times (0, T]$. In view of Lemma 2, $\delta_\theta(z_1, z_2, \tau)$ is a viscosity solution, and therefore a sub-solution of

$$\partial_\tau f + H_\theta(z_1, \tau, \partial_{z_1} f) - H_\theta(z_2, \tau, -\partial_{z_2} f) = 0, \text{ on } \Upsilon. \tag{D.3}$$

We aim to find an appropriate non-negative function $\phi(z_1, z_2, \tau)$ independent of θ such that ϕ is (1) a super-solution of (D.3) - i.e a super-solution of (D.3) for all $\theta \in \Theta$ - on some convenient domain $\Omega \equiv \mathcal{A} \times (0, T]$, where $\mathcal{A} \subseteq (\underline{z}, \infty) \times (\underline{z}, \infty)$, and (2) that also satisfies $\phi \geq \delta_\theta$ on $\Gamma \equiv \{\partial\mathcal{A} \times (0, T]\} \cup \{\bar{\mathcal{A}} \times \{0\}\}$ - again for all $\theta \in \Theta$. Then by the Comparison Theorem 5, we will be able to obtain $\delta_\theta \leq \phi$ on $\bar{\Omega}$.[43] We claim that such a function is given by

$$\phi(z_1, z_2, \tau) := Ae^{B\tau}\left(|z_1 - z_2|^2 + \varepsilon\right)^{1/2}$$

after choosing $\mathcal{A} := \{(z_1, z_2) : |z_1 - z_2| < 1, \underline{z} < z_1, \underline{z} < z_2\}$. Here, A, B are some appropriately chosen constants and $\varepsilon > 0$ is an arbitrarily small number (we shall later send this to 0).[44]

First note that under the choice of set $\mathcal{A}$, ϕ is continuous and uniformly bounded, as demanded by the definition of a viscosity super-solution.

Next, let us show that for all $\theta \in \Theta$, $\phi \geq \delta_\theta$ on $\Gamma \equiv \{\partial\mathcal{A} \times (0, T]\} \cup \{\bar{\mathcal{A}} \times \{0\}\}$, under some appropriate choice of A. Clearly, $\phi \geq \delta_\theta$ on $\bar{\mathcal{A}} \times \{0\}$ since $\phi(z_1, z_2, 0) \geq 0$ for all (z_1, z_2), while $\delta_\theta(z_1, z_2, 0) = 0$. It therefore remains to show $\phi \geq \delta_\theta$ on $\partial\mathcal{A} \times (0, T]$. We have three (not necessarily exclusive) possibilities for $\partial\mathcal{A}$: (i) $|z_1 - z_2| = 1$, (ii) $z_1 = \underline{z}$, or (iii) $z_2 = \underline{z}$. In the first case, i.e when $|z_1 - z_2| = 1$, we have $\phi(z_1, z_2, \tau) \geq e^{B\tau}A$. Now, by Lemma 3, $|u_\theta| \leq K$ for some $K < \infty$ independent of θ. Hence, as long as we choose $A \geq 2K$, we can ensure $\phi \geq \delta_\theta$ on the region of $\partial\mathcal{A}$ where $|z_1 - z_2| = 1$. Next, consider the case when $z_1 = \underline{z}$. Here $\phi(\underline{z}, z_2, \tau) \geq e^{B\tau}A(z_2 - \underline{z})$. But $u_\theta(\underline{z}, \tau) = 0$, while by Lemma 3, $u_\theta(z_2, \tau) \leq K_1(z_2 - \underline{z})$, where $K_1 < \infty$ is independent of θ, τ. Thus here too we can ensure $\phi \geq \delta_\theta$ by choosing $A \geq K_1$. A symmetric argument also implies $\phi \geq \delta_\theta$ when $z_2 = \underline{z}$. In view of the above, we have thus shown that there exists $A < \infty$ for which $\phi \geq \delta_\theta$ on Γ.

[43] Note that the Comparison theorem is now being applied on (D.3). Let $\mathbf{z} = (z_1, z_2)^\intercal$ and $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2)^\intercal$. Then it is straightforward to verify that the Hamiltonian $\tilde{H}_\theta(\mathbf{z}, t, \mathbf{p}) := H_\theta(z_1, \tau, \mathbf{p}_1) - H_\theta(z_2, \tau, \mathbf{p}_2)$ satisfies the properties (H1)-(H3) in view of Assumption 1.

[44] The reason for not setting $\varepsilon = 0$ straightaway is to ensure $\left(|z_1 - z_2|^2 + \varepsilon\right)^{1/2}$ is differentiable everywhere.

We now show that for all $\theta \in \Theta$, ϕ is a super-solution of (D.3) on the domain Ω, under some appropriate choice of B (given A). To this end, observe that

$$
\begin{aligned}
&\partial_\tau \phi + H_\theta\left(z_1, \tau, \partial_{z_1}\phi\right) - H_\theta\left(z_2, \tau, -\partial_{z_2}\phi\right) \\
&= ABe^{B\tau}\left(|z_1 - z_2|^2 + \varepsilon\right)^{1/2} \\
&\quad + H_\theta\left(\tau, z_1, \frac{Ae^{B\tau}(z_1 - z_2)}{(|z_1 - z_2|^2 + \varepsilon)^{1/2}}\right) - H_\theta\left(\tau, z_2, \frac{Ae^{B\tau}(z_1 - z_2)}{(|z_1 - z_2|^2 + \varepsilon)^{1/2}}\right) \\
&:= ABe^{B\tau}\left(|z_1 - z_2|^2 + \varepsilon\right)^{1/2} + \Delta_\theta(\tau, z_1, z_2; A, B).
\end{aligned} \tag{D.4}
$$

Now under Assumptions 1(i)-(ii) -which ensures $\bar{G}_\theta(z,t)$ and $\bar{r}_\theta(z,t)$ are uniformly Lipschitz continuous - and some straightforward algebra, we have

$$
\begin{aligned}
|\Delta_\theta(\tau, z_1, z_2; A, B)| &\leq Ae^{B\tau}\lambda(\tau)\left|\bar{G}_\theta(z_1, \tau) - \bar{G}_\theta(z_2, \tau)\right| + e^{\beta\tau}\lambda(\tau)\left|\bar{r}_\theta(z_1, \tau) - \bar{r}_\theta(z_2, \tau)\right| \\
&\leq Ae^{\max\{B,\beta\}\tau}\lambda(\tau)M\left|z_1 - z_2\right|,
\end{aligned}
$$

for some constant $M < \infty$ independent of θ, z_1, z_2, τ. Plugging the above expression into (D.4), we note that by choosing B large enough (e.g $B \geq \max\{AM\bar{\lambda}, \beta\}$ would suffice), it follows

$$
\partial_\tau \phi + H_\theta\left(\tau, z_1, \partial_{z_1}\phi\right) - H_\theta\left(\tau, z_2, -\partial_{z_2}\phi\right) \geq 0 \text{ on } \Omega,
$$

for all $\theta \in \Theta$. This implies that for all $\theta \in \Theta$, ϕ is a super-solution of (D.3) on Ω.

By now we shown that for all $\theta \in \Theta$, $\phi \geq \delta_\theta$ on Γ, and that ϕ is a super-solution of (D.3) on Ω. At the same time, δ_θ is viscosity sub-solution of (D.3) on Ω. Hence by applying the Comparison theorem on (D.3), we get $\phi \geq \delta_\theta$ on $\bar{\Omega}$, i.e

$$
u_\theta(z_1, \tau) - u_\theta(z_2, \tau) \leq e^{B\tau}\left(A|z_1 - z_2|^2 + \varepsilon\right)^{1/2}
$$

for all $(z_1, z_2, \tau) \in \bar{\Omega}$ and $\theta \in \Theta$. But the choice of ε was arbitrary. We may therefore take this to 0 to obtain

$$
\sup_{(z_1, z_2, \tau) \in \bar{\Omega}, \theta \in \Theta}\left(u_\theta(z_1, \tau) - u_\theta(z_2, \tau) - Ae^{B\tau}|z_1 - z_2|\right) \leq 0
$$

Now, $\bar{\Omega} \equiv \bar{\mathcal{A}} \times [0, T]$, where $\bar{\mathcal{A}}$ includes all z_1, z_2 such that $|z_1 - z_2| < 1$. Hence, we can conclude that $u_\theta(\cdot, t)$ is locally Lipschitz in its first argument. Since $h_\theta(\cdot, t) = e^{\beta(T-t)}u_\theta(\cdot, T - t)$, this also implies that h_θ is locally Lipschitz continuous in its first argument. $\square$

D.2. **Periodic boundary condition.** Following (3.5), we consider time periodic first order PDEs of the form

$$
\begin{aligned}
&\partial_t f + H\left(z, t, f(z,t), \partial_z f(z,t)\right) = 0 \text{ on } \mathcal{U}; \\
&\qquad\qquad f(z,t) = f(z, t + T_p)\ \forall (z,t) \in \mathcal{U}.
\end{aligned} \tag{D.5}
$$

We first present a stronger version of the Comparison Theorem for Cauchy problems, due to Crandall and Lions (1983). This turns out to be useful to prove a Comparison theorem for periodic problems as in Bostan and Namah (2007). Denote $(f)_+ := \max\{f, 0\}$.

Lemma 5. *Suppose that the function $H(\cdot)$ satisfies conditions (H1)-(H3). Let u, v be, respectively, viscosity sub- and super-solutions to*

$$\partial_t f + H\left(z, t, f(z,t), \partial_z f(z,t)\right) = 0 \text{ on } \mathbb{R} \times (t_0, T].$$

Then for all $t \in [t_0, T]$,

$$e^{\beta(t-t_0)} \sup_{z \in \mathbb{R}^d} \left(u(z,t) - v(z,t)\right)_+ \leq \sup_{z \in \mathbb{R}^d} \left(u(z,t_0) - v(z,t_0)\right)_+ .$$

Theorem 6. ***(Comparison Theorem - Periodic form)*** *Suppose that the function $H(\cdot)$ satisfies conditions (H1)-(H3), and that it is T_p-periodic in t. Let u, v be respectively, T_p-periodic viscosity sub- and super-solutions to (D.5) on $\mathcal{U}$. Then $u(x,t) \leq v(x,t)$ on $\mathbb{R} \times \mathbb{R}$.*

Proof. By Lemma 5, we have that for any $t_0 \in \mathbb{R}$,

$$e^{\beta T_p} \sup_{z \in \mathbb{R}} \left(u(z, T_p + t_0) - v(z, T_p + t_0)\right)_+ \leq \sup_{z \in \mathbb{R}} \left(u(z,t_0) - v(z,t_0)\right)_+ .$$

But by periodicity, $u(z, T_p + t_0) - v(z, T_p + t_0) = u(z, t_0) - v(z, t_0)$, hence it must be the case $\sup_{z \in \mathbb{R}} \left(u(z,t_0) - v(z,t_0)\right)_+ = 0$. But the choice of t_0 was arbitrary; therefore $u(z,t) \leq v(z,t)$ on $\mathbb{R} \times \mathbb{R}$. □

Lemma 6. *Suppose that Assumptions 1-4 hold for the periodic boundary condition, and the discount factor β is sufficiently large. Then there exists $L_1 < \infty$ independent of θ, z, t such that h_θ is locally Lipschitz continuous with Lipschitz constant L_1.*

Proof. We first show that $h_\theta(\cdot, t)$ is Lipschitz continuous in its first argument. As in the proof of Theorem 2, fix any $t^* > T_p$, and denote $u_\theta(z, \tau) := e^{\beta\tau} h_\theta(z, t^* - \tau)$. Also, let $\delta_\theta(z_1, z_2, \tau) := u_\theta(z_1, \tau) - u_\theta(z_2, \tau)$ and recall that

$$H_\theta(z, \tau, p) := -e^{\beta\tau}\lambda(\tau)\bar{r}_\theta(z,\tau) - \lambda(\tau)\bar{G}_\theta(z,\tau)p.$$

In view of Lemma 2, $\delta_\theta(z_1, z_2, \tau)$ is a viscosity solution, and therefore a sub-solution of

$$\partial_\tau f + H_\theta\left(\tau, z_1, \partial_{z_1} f\right) - H_\theta\left(\tau, z_2, -\partial_{z_2} f\right) = 0, \text{ on } \Omega, \tag{D.6}$$

where

$$\Omega \equiv \mathcal{A} \times (0, T_p]; \; \mathcal{A} \equiv \{(z_1, z_2) : |z_1 - z_2| < 1\}.$$

We shall compare δ_θ against the function

$$\phi(z_1, z_2, \tau) := A e^{B\tau} \left(|z_1 - z_2|^2 + \varepsilon\right)^{1/2}.$$

By the same arguments as in the proof of Lemma 4, we can set $B = \beta$ and choose A in such a way that $\phi \geq \delta_\theta$ on $\partial\mathcal{A} \times (0, T_p]$, and ϕ is a super-solution to (D.6). This step requires β to be sufficiently large ($\beta \geq AM\bar{\lambda}$ would suffice), as assumed in the statement of Theorem 2. Subsequently, by the Comparison Theorem 5, we obtain

$$\sup_{z_1, z_2 \in \mathbb{R}^2} \left(u_\theta(z_1, T_p) - u_\theta(z_2, T_p) - \phi(z_1, z_2, T_p)\right)_+ \leq \sup_{z_1, z_2 \in \mathbb{R}^2} \left(u_\theta(z_1, 0) - u_\theta(z_2, 0) - \phi(z_1, z_2, 0)\right)_+ .$$

Rewriting the above in terms of h_θ, and noting that $h_\theta(z,\cdot)$ is T_p-periodic, we get

$$e^{\beta T_p} \sup_{z_1,z_2\in\mathbb{R}^2} \left(h_\theta(z_1,t^*) - h_\theta(z_2,t^*) - e^{-\beta T_p}\phi(z_1,z_2,T_p)\right)_+ \le \sup_{z_1,z_2\in\mathbb{R}^2} \left(h_\theta(z_1,t^*) - h_\theta(z_2,t^*) - \phi(z_1,z_2,0)\right)_+ .$$

Since we set $B=\beta$, we have $e^{-\beta T_p}\phi(z_1,z_2,T_p) = \phi(z_1,z_2,0)$. In view of the above,

$$\sup_{z_1,z_2\in\mathbb{R}^2} \left(h_\theta(z_1,t^*) - h_\theta(z_2,t^*) - \phi(z_1,z_2,0)\right)_+ \le 0.$$

Since t^* is arbitrary, this proves the Lipschitz continuity of h_θ with respect to z, after sending $\varepsilon \to 0$ in the definition of ϕ.

We now show that $h_\theta(z,\cdot)$ is Lipschitz continuous in its second argument. For this, we will use the original form of the PDE. Rewrite PDE (3.2) in the form $\partial_t f + \bar{H}_\theta(z,t,f,\partial_z f) = 0$, where

$$\bar{H}_\theta(z,t,u,p) := -\beta u + \lambda(t)\bar{r}_\theta(z,t) + \lambda(t)\bar{G}_\theta(z,t)p,$$

and consider the Cauchy problem

$$\text{(D.7)} \qquad \begin{aligned} \partial_t f + \bar{H}_\theta(z,t,f,\partial_z f) &= 0 \text{ on } \mathbb{R}\times(t_1,\infty); \\ f(\cdot,t_1) &= v_0, \end{aligned}$$

for any continuous function v_0. Denote the solution of the above as f_θ. We now compare f_θ with $\phi := v_0 + K(t-t_1)$, for some constant K. Indeed, arguing as in the proof of Lemma 3, we can find $K<\infty$ independent of θ, z, t, t_1 such that ϕ is a viscosity super-solution of $\partial_t f + \bar{H}_\theta(z,t,f,\partial_z f) = 0$ on $\mathbb{R}\times(t_1,\infty)$. Also, $\phi = v_0 = f_\theta$ on $\mathbb{R}\times\{t_1\}$. Hence, by the Comparison Theorem 5, $\phi \ge f_\theta$ on $\mathbb{R}\times[t_1,\infty)$, i.e $f_\theta - v_0 \le K(t-t_1)$. A symmetric argument involving $\varphi := v_0 - K(t-t_1)$ as a sub-solution will also show that $v_0 - f_\theta \le K(t-t_1)$. Taken together, we obtain

$$\sup_{z\in\mathbb{R}} |f_\theta(z,t) - v_0(z)| \le K(t-t_1).$$

Note that this inequality holds uniformly over all continuous v_0. In particular, we may set $v_0(\cdot) = h_\theta(\cdot,t_1)$. But with this initial condition the unique solution of (D.7) on $\mathbb{R}\times[t_1,\infty)$ is simply h_θ itself (i.e $f_\theta \equiv h_\theta$). We have thus shown that $\sup_{z\in\mathbb{R}} |h_\theta(z,t) - h_0(z,t_1)| \le K(t-t_1)$ for all $t \ge t_1$. But the choice of t_1 here was arbitrary. Consequently, this property holds for all $t_1 \in \mathbb{R}$ and $t \ge t_1$, which implies that $h_\theta(z,\cdot)$ is Lipschitz continuous in its second argument uniformly over θ, z. □

D.3. **Neumann and Periodic-Neumann boundary conditions.** For results on the Neumann and periodic-Neumann boundary conditions, we go back to the more general version of first order PDEs as given in (A.2). We shall impose a couple of additional regularity conditions on $F(\cdot)$ and $B(\cdot)$, in addition to (R1)-(R8) in Appendix A. These are given by (as before, we use the notation $y := (z,t)$):

(R9) There exist $C_1, C_2 < \infty$ such that

$$\begin{aligned} |F(y_1,u,p_1) - F(y_2,u,p_2)| &\le C_1\left(\|y_1-y_2\| + \|p_1-p_2\|\right), \quad \text{and} \\ |F(y_1,u,p) - F(y_2,u,p)| &\le C_2\|p\|\,\|y_1-y_2\|. \end{aligned}$$

(R10) There exist $C_3, C_4 < \infty$ such that

$$|B(y_1, u, p_1) - B(y_2, u, p_2)| \leq C_3 (\|y_1 - y_2\| + \|p_1 - p_2\|), \quad \text{and}$$
$$|B(y_1, u, p) - B(y_2, u, p)| \leq C_4 \|p\| \|y_1 - y_2\|.$$

It is straightforward to verify that all the regularity conditions (R1)-R(10) are satisfied for our population PDE (3.2) under Assumption 1, with constants C_1, C_2, C_3, C_4 independent of θ (this is due to uniform Lipschitz continuity of $\bar{G}_\theta(z,t)$ and $\bar{r}_\theta(z,t)$ imposed in Assumption 1(i)). It is useful to note that for the population PDE, the first part of Assumption 1(i) is not required to verify these conditions. On a side note, we also remark that the sample PDE (3.10) may not satisfy the regularity conditions (R9) and (R10). This is because Assumption 1 does not imply $\hat{G}_\theta(z,t)$ and $\hat{r}_\theta(z,t)$ are Lipschitz continuous, only that they are continuous (where the continuity is turn due to the first part of Assumption 1(i)).

The following results are taken from Barles & Lions (1991), but see also Crandall, Ishii & Lions (1992, Theorem 7.12). We refer to those papers for the proofs.

Theorem 7. ***(Comparison Theorem - Neumann form)*** *Suppose that the functions $F(\cdot)$ and $B(\cdot)$ satisfies conditions (R1)-(R8) in Appendix A. Let u, v be respectively, a viscosity sub- and super-solutions to (A.2). Then $u(x,t) \leq v(x,t)$ on $\bar{\mathcal{Z}} \times [0, \bar{T}]$.*

Lemma 7. *Suppose that the functions $F(\cdot)$ and $B(\cdot)$ satisfies conditions (R1)-(R10). Then the unique viscosity solution, u, to (A.2) is Lipschitz continuous on $\bar{\mathcal{Z}} \times [0, \bar{T}]$, where the Lipschitz constant depends only on the values of C_1-C_4 and M in (R1)-(R10).*

The next set of results are for the periodic-Neumann boundary condition. These follow from Theorem 7 and Lemma 7 in the same way that Theorem 6 and Lemma 6 follow from Theorem 5 and Lemma 4, and are therefore also presented without a proof.

Theorem 8. ***(Comparison Theorem - Periodic Neumann form)*** *Suppose that the functions $F(\cdot)$ and $B(\cdot)$ satisfies conditions (R1)-(R8) in Appendix A, and that they are both also T_p-periodic in t. Let u, v be respectively, T_p-periodic viscosity sub- and super-solutions to (A.2). Then $u(x,t) \leq v(x,t)$ on $\bar{\mathcal{Z}} \times \mathbb{R}$.*

Lemma 8. *Suppose that the functions $F(\cdot)$ and $B(\cdot)$ satisfies conditions (R1)-(R10), they are both also T_p-periodic, and the discount factor β is sufficiently large. Then the unique T_p-periodic viscosity solution, u, to (A.2) is Lipschitz continuous on $\bar{\mathcal{Z}} \times \mathbb{R}$, where the Lipschitz constant depends only on the values of T_p and C_1-C_4 & M in (R1)-(R10).*

Appendix E. Semi-convexity, sup-convolution etc.

In this Section, we collect various properties of semi-convex/concave functions, and sup/inf-convolutions used in the proof of Theorem 3. Many of these results are well known. In some cases we provide simpler proofs at the expense of obtaining results that are not as sharp, but they will suffice for the purpose of proving the theorems in this paper.

E.1. **Semi-convexity and concavity.**

In what follows we take y to be a vector in $\mathbb{R}^n$. Also, for any vector y, $|y|$ denotes its Euclidean norm.

Definition 3. *A function u on $\mathbb{R}^n$ is said to be semi-convex with the coefficient c if $u(y) + \frac{c}{2}|y|^2$ is a convex function. Similarly, u is said to be semi-concave with the coefficient c if $u(y) - \frac{c}{2}|y|^2$ is concave.*

The following lemma states a useful property of semi-convex functions.

Lemma 9. *Suppose that u is semi-convex. Then u is twice differentiable almost everywhere. Furthermore, for every point at which Du exists, we have for all $h \in \mathbb{R}^n$,*

$$u(y+h) \geq u(y) + h^\intercal Du(y) - \frac{c}{2}|h|^2.$$

Proof. Define $g(y) = u(y) + \frac{c}{2}|y|^2$. Since $g(y)$ is convex, the Alexandrov theorem implies $g(\cdot)$ is twice continuously differentiable almost everywhere. Hence $u(y) = g(y) - \frac{c}{2}|y|^2$ is also twice differentiable almost everywhere.

For the second part of the theorem, observe that by convexity,

$$g(y+h) \geq g(y) + h^\intercal Dg(y).$$

Note that where the derivative exists, $Dg(y) = Du(y) + cy$. Hence,

$$u(y+h) + \frac{c}{2}|y+h|^2 \geq u(y) + \frac{c}{2}|y|^2 + h^\intercal Du(y) + ch^\intercal y.$$

Rearranging the above expression gives the desired inequality. □

An analogous property also holds for semi-concave functions. We can also extend the scope of the theorem to points where Du does not exist by considering one-sided derivatives, which can be shown to exist everywhere for semi-convex functions.

E.2. **Sup and Inf Convolutions.**

Let $u(y)$ denote a continuous function on some open set $\mathcal{Y}$. Let $\partial\mathcal{Y}$ denote the boundary of $\mathcal{Y}$, and $\bar{\mathcal{Y}}$ its closure. Additionally, for some function u, we let $\|Du\|$ denote the Lipschitz constant for u, with the convention that it is ∞ if u is not Lipschitz continuous.

Definition 4. *The function u^ϵ is said to be the sup-convolution of u if*

$$u^\epsilon(y) = \sup_{w \in \bar{\mathcal{Y}}} \left\{ u(w) - \frac{1}{2\epsilon}|w-y|^2 \right\}.$$

Similarly, u_ϵ is said to be the inf-convolution of u if

$$u_\epsilon(y) = \inf_{w \in \bar{\mathcal{Y}}} \left\{ u(w) + \frac{1}{2\epsilon}|w-y|^2 \right\}.$$

We shall also define y^ϵ as the value for which

$$u(y^\epsilon) - \frac{1}{2\epsilon}|y^\epsilon - y|^2 = u^\epsilon(y),$$

if y^ϵ lies in $\mathcal{Y}$ (otherwise it is taken to be undefined). Analogously, y_ϵ is the value for which

$$u(y_\epsilon) + \frac{1}{2\epsilon}|y_\epsilon - y|^2 = u_\epsilon(y).$$

Additionally, define $\mathcal{Y}^\epsilon$ as the set of all points in $\mathcal{Y}$ that are atleast $2\|Du\|\epsilon$ distance away from $\partial\mathcal{Y}$, i.e

$$\mathcal{Y}^\epsilon := \{y \in \mathcal{Y} : |y - w| \geq 2\|Du\|\epsilon \ \forall w \in \partial\mathcal{Y}\}.$$

We have the following properties for sup and inf-convolutions:

Lemma 10. *Suppose that u is continuous. Then,*

(i) u^ϵ is semi-convex with coefficient $1/\epsilon$. Similarly, u_ϵ is semi-concave with coefficient $1/\epsilon$.

(ii) $|y^\epsilon - y| \leq 2\|Du\|\epsilon$ and $|y_\epsilon - y| \leq 2\|Du\|\epsilon$.

(iii) For all $y \in \mathcal{Y}^\epsilon$, $|u^\epsilon(y) - u(y)| \leq 4\|Du\|^2\epsilon$ and $|u_\epsilon(y) - u(y)| \leq 4\|Du\|^2\epsilon$.

Proof. We show the above properties for u^ϵ and y^ϵ. The claims for u_ϵ and y_ϵ follow in an analogous manner.

For (i), observe that

$$u^\epsilon(y) + \frac{1}{2\epsilon}|y|^2 = \sup_{w\in\bar{\mathcal{Y}}}\left\{u(w) + \frac{1}{\epsilon}w^\intercal y - \frac{1}{2\epsilon}|w|^2\right\}.$$

The right hand side of the above expression is in the form of a supremum over affine functions, which is convex. Hence (i) follows by the definition of semi-convex functions.

For (ii), by the definition of y^ϵ and u^ϵ,

$$\frac{1}{2\epsilon}|y^\epsilon - y|^2 \leq u(y^\epsilon) - u(y) \leq \|Du\|\,|y^\epsilon - y|.$$

Rearranging the above inequality we get $|y^\epsilon - y| \leq 2\|Du\|\epsilon$.

For (iii), by the definition of y^ϵ (which exists for $y \in \mathcal{Y}^\epsilon$ in view of part (ii)),

$$\begin{aligned} |u^\epsilon(y) - u(y)| &= \left|u(y^\epsilon) - u(y) + \frac{1}{2\epsilon}|y^\epsilon - y|^2\right| \\ &\leq \|Du\|\,|y^\epsilon - y| + \frac{1}{2\epsilon}|y^\epsilon - y|^2 \\ &\leq 4\|Du\|^2\epsilon, \end{aligned}$$

where the last inequality follows by (ii). □

Lemma 11. *Assume that u is uniformly continuous. Suppose that $\phi \in C^2(\mathcal{Y})$, such that $u^\epsilon - \phi$ has a local maximum at $y_0 \in \mathcal{Y}^\epsilon$. Define $\psi(y) = \phi(y + y_0 - y_0^\epsilon)$. Then $u - \psi$ has a local maximum at $y_0^\epsilon \in \mathcal{Y}$, and*

$$D\psi(y_0^\epsilon) = D\phi(y_0) = \frac{1}{\epsilon}(y_0^\epsilon - y_0).$$

Proof. Since $u^\epsilon - \phi$ has a local maximum at y_0, this implies there is a ball $B(y_0, r)$ of radius r around y_0 for which

$$u^\epsilon(y_0) - \phi(y_0) \geq u^\epsilon(w) - \phi(w)$$

for all $w \in B(y_0, r)$. Hence,

$$u(y_0^\epsilon) - \frac{1}{2\epsilon}|y_0^\epsilon - y_0|^2 - \phi(y_0) \geq u^\epsilon(w) - \phi(w)$$
$$\geq u(y) - \frac{1}{2\epsilon}|w - y|^2 - \phi(w)$$

for all y and $w \in B(y_0, r)$ (note that $y_0^\epsilon \in \mathcal{Y}$ in view of the definition of $\mathcal{Y}^\epsilon$ and Lemma 10). This implies that (y_0^ϵ, y_0) is the local maximum of the function

$$\Upsilon(y, w) := u(y) - \frac{1}{2\epsilon}|w - y|^2 - \phi(w).$$

In other words,

$$\Upsilon(y_0^\epsilon, y_0) \geq \Upsilon(y, w) \quad \forall y \text{ and } w \in B(y_0, r). \tag{E.1}$$

In view of (E.1), we have $\Upsilon(y_0^\epsilon, y_0) \geq \Upsilon(w - y_0 + y_0^\epsilon, w)$ for all $w \in B(y_0, r)$, which implies

$$u(y_0^\epsilon) - \frac{1}{2\epsilon}|y_0^\epsilon - y_0|^2 - \phi(y_0) \geq u(w - y_0 + y_0^\epsilon) - \frac{1}{2\epsilon}|y_0^\epsilon - y_0|^2 - \phi(w).$$

Hence, for all $w \in B(y_0, r)$,

$$u(y_0^\epsilon) - \phi(y_0) \geq u(w - y_0^\epsilon + y^\epsilon) - \phi(w).$$

Now set $y^* = w - y_0 + y_0^\epsilon$ and observe that $|y^* - y_0^\epsilon| = |w - y_0| \leq r$ for all $w \in B(y_0, r)$. We thus obtain that for all $y^* \in B(y^\epsilon, r)$,

$$u(y_0^\epsilon) - \phi(y_0) \geq u(y^*) - \phi(y^* + y_0 - y_0^\epsilon).$$

In view of the definition of $\psi(.)$, the above implies

$$u(y_0^\epsilon) - \psi(y_0^\epsilon) \geq u(y^*) - \psi(y^*) \ \forall y^* \in B(y_0^\epsilon, r).$$

Hence $u - \psi$ has a local maximum at y_0^ϵ.

For the second part of the lemma, observe that by (E.1), $\Upsilon(y_0^\epsilon, y_0) \geq \Upsilon(y_0^\epsilon, w)$ for all $w \in B(y_0, r)$, which implies (after some rearrangement)

$$\frac{1}{2\epsilon}|y_0^\epsilon - w|^2 + \phi(w) \geq \frac{1}{2\epsilon}|y_0^\epsilon - y_0|^2 + \phi(y_0), \ \forall w \in B(y_0, r).$$

Hence the function $\theta(w) := \frac{1}{2\epsilon}|y_0^\epsilon - w|^2 + \phi(w)$ has a local minimum at $w = y_0$. Consequently,

$$D\phi(y_0) = \frac{1}{\epsilon}(y_0^\epsilon - y_0).$$

This proves the second claim after noting $D\psi(y_0^\epsilon) = D\phi(y_0)$. □

Our next Lemma considers PDEs of the form

$$F(y, u(y), Du(y)) = 0 \text{ on } \mathcal{Y}.$$

We shall assume that $F(\cdot)$ satisfies the following property (here $C < \infty$ denotes some constant)

$$|F(y_1, q_1, p) - F(y_2, q_2, p)| \leq Cp\{|q_1 - q_2| + |y_1 - y_2|\}. \tag{E.2}$$

Lemma 12. *Suppose that u is a viscosity solution of $F(y, u, Du) = 0$, and $\|Du\| \leq m < \infty$. Suppose also that $F(\cdot)$ satisfies (E.2). Then there exists some c depending on only C (from E.2) and m such that for all $y \in \mathcal{Y}^\epsilon$,*

$$F(y, u^\epsilon, Du^\epsilon) \leq c\epsilon,$$

where the above holds in the viscosity sense.

Proof. Take any $\phi \in C^2(\mathcal{Y})$ such that $u^\epsilon - \phi$ has a local maximum at $y_0 \in \mathcal{Y}^\epsilon$. Set $\psi(y) := \phi(y + y_0 - y_0^\epsilon)$. Then by Lemma 11, $u - \psi$ has a local maximum at $y_0^\epsilon \in \mathcal{Y}$. Hence, by definition of the viscosity solution

$$F(y, u(y_0^\epsilon), D\psi(y_0^\epsilon)) \leq 0. \tag{E.3}$$

Recall also from Lemma 11 that

$$|D\psi(y_0^\epsilon)| = |D\phi(y_0)| = \frac{1}{\epsilon}|y_0^\epsilon - y_0| < \|Du\| \leq m.$$

We then have

$$\begin{aligned}
&|F(y_0, u^\epsilon(y_0), D\phi(y_0)) - F(y_0^\epsilon, u(y_0^\epsilon), D\psi(y_0^\epsilon))| \\
&\leq Cm \{|y_0 - y_0^\epsilon| + |u^\epsilon(y_0) - u(y_0^\epsilon)|\} \\
&\leq Cm \{(1 + m)|y_0 - y_0^\epsilon| + |u^\epsilon(y_0) - u(y_0)|\} \\
&\leq Cm\{2m(1 + m) + 4m^2\}\epsilon := c\epsilon,
\end{aligned}$$

where the first inequality follows from (E.2) and the last inequality from Lemma 10. We thus obtain in view of the above and (E.3) that

$$F(y_0, u^\epsilon(y_0), D\phi(y_0)) \leq c\epsilon. \tag{E.4}$$

Since c is a constant, we have thus shown that if $u^\epsilon - \phi$ has a local maximum at some $y_0 \in \mathcal{Y}^\epsilon$, then (E.4) holds. This implies that in a viscosity sense

$$F(y, u^\epsilon, Du^\epsilon) \leq c\epsilon.$$

□

Appendix F. Estimation of clusters and cluster specific arrival rates

Section 6.2 in the main text discussed a setup in which the distribution of the arrivals varied with time. In this section we discuss how both the clusters and the cluster specific arrival rates can be estimated. We shall suppose, as in our empirical application, that we have access to a time series dataset consisting of a sample of $(N_d, \mathbf{x}_d)_{d=1}^{D}$, where N_d denotes the number of people of people that arrived on day d, and $\mathbf{x}_d$ denotes the covariates of all the individuals who arrived on that day. Note that d denotes the unit of aggregation in the data. These observations can generally come from a different dataset than the observational one used to estimate the rewards.

First, let us suppose that there is no heterogeneity in the arrival rates between individuals. Then we can parameterize $\lambda(t) \approx \exp\left\{\sum_{l=1}^{k} \beta_l \vartheta_l(t)\right\}$, where $\{\vartheta_l(\cdot)\}_{l=1}^{\infty}$ is a set of basis functions, e.g Fourier series. Let $t(d)$ denote the time corresponding to day d (recall that time is rescaled). Then the coefficients $\beta := (\beta_1, \dots, \beta_k)$ can be estimated by maximizing the likelihood

$$l(\mathbf{N}|\beta) := \prod_{d=1}^{D} \frac{\lambda(t;\beta)^{N_d}}{N_d!} e^{-\lambda_k(t;\beta)}, \quad \text{where } \lambda_k(t;\beta) := \exp\left\{\sum_{l=1}^{k} \beta_l \vartheta_l(t)\right\}.$$

Letting $\hat{\beta}$ denote the MLE estimates, the estimate of arrival rates is given by

$$\hat{\lambda}(t) = \exp\left\{\sum_{l=1}^{k} \hat{\beta}_l \vartheta_l(t)\right\}.$$

We now turn to estimation when there is heterogeneity in the arrival rates. As we mentioned in the main text, we can handle the heterogeneity using clustering. The basic premise is to partition the data into blocks $c = \{1, \dots, J\}$ and estimate $\hat{\lambda}_c(t)$ separately for each cluster. This a two step procedure. There are many ways one could do the first step; in this paper we recommend k-mean or k-median clustering due to its simplicity and ease of use.

In our empirical application, we employ k-median clustering (a well-established method, for full details see Anderberg, 1973). The aim is to divide the candidates into k clusters. The clusters are chosen such that the characteristics of each candidate are as close as possible to the characteristic-medians of their cluster. The clusters are chosen such that the squared sum of Euclidean distances between the vector of characteristics of each candidate and the vector of characteristics-medians of their cluster is as small as possible.

In practice, we use Lloyd's algorithm, as usual for k-median clustering. First, we start with k randomly selected candidates (k can be chosen freely, $k = 4$ in our JTPA example), which are each the 'founding members' of each cluster. All other candidates that are then allocated to the cluster with the smallest Euclidean distance between the vector of characteristics of the candidate and the founding member. Second, the median of each cluster's characteristics is computed and denoted 'centroid'. Each candidate is then re-allocated to the cluster with the smallest Euclidean distance between the vector of characteristics of the candidate and the centroid. The second step is repeated until convergence, i.e until no more re-allocations occur.

Appendix G. JTPA Application: Additional Figures and Tables

G.1. **Clusters.** Table 1 describes the clusters resulting from the JTPA example. Cluster 1 appears to contain predominantly candidates with high previous earnings. Cluster 2's distinguishing factor is the high age, and for cluster 3 it is few years of education. Cluster 4 contains young educated candidates with low previous earnings.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Age: Mean	31.8	44.9	31.3	26.9
Age: Min.	22	34	22	22
Age: Max.	63	78	57	34
Prev. Earnings: Mean	8999	1439	1413	1231
Prev. Earnings: Min.	3600	0	0	0
Prev. Earnings: Max.	63000	12000	9076	5130
Education: Mean	12.1	12.1	9.0	12.3
Education: Min.	7	8	7	11
Education: Max.	18	18	10	18
Observations	2278	2198	1698	3049

Table 1. Cluster Descriptions

G.2. **Results & Interpretation.** In order to further interpret the resulting (final) policy function, we use that function in 100 evaluation episodes and record the treated candidates. As a measure of selectivity, we record how many candidates were declined before one was treated. Figure G.1 illustrates for each treated person in the 100 episodes how many candidates were declined since the last treatment, plotted against the remaining budget. In the both cases, the algorithm becomes more selective when the budget is scarce. This is in line with economic considerations due to discounting. The inter-temporal trade-off is between treating a person now versus treating a person at the end of the budget. If the remaining budget is large, the reward from treating a person at the end of the budget is discounted more heavily compared to the case where the remaining budget is small. There are 160,000 points depicted in Figure G.1 and outliers appear over-prominent. Figure G.2 simplifies this illustration by imposing a linear/quadratic structure.

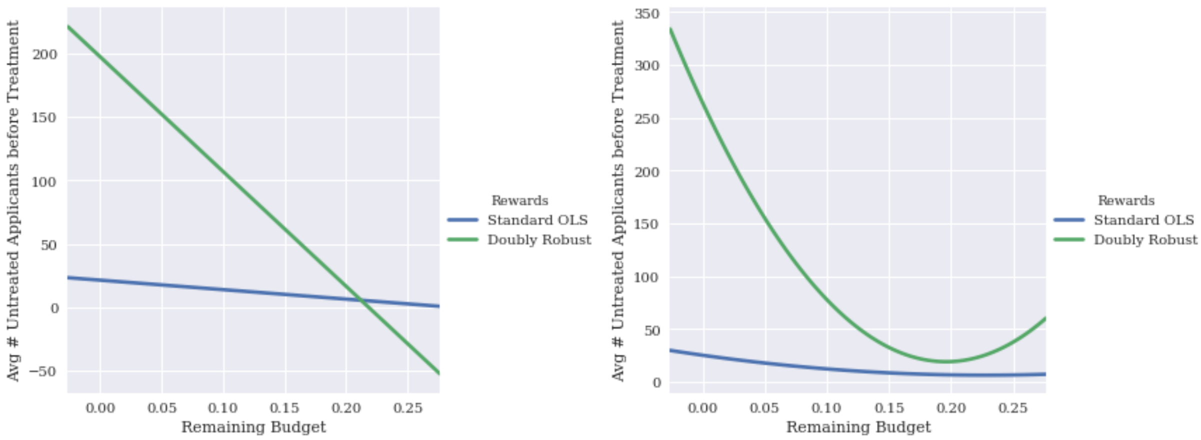
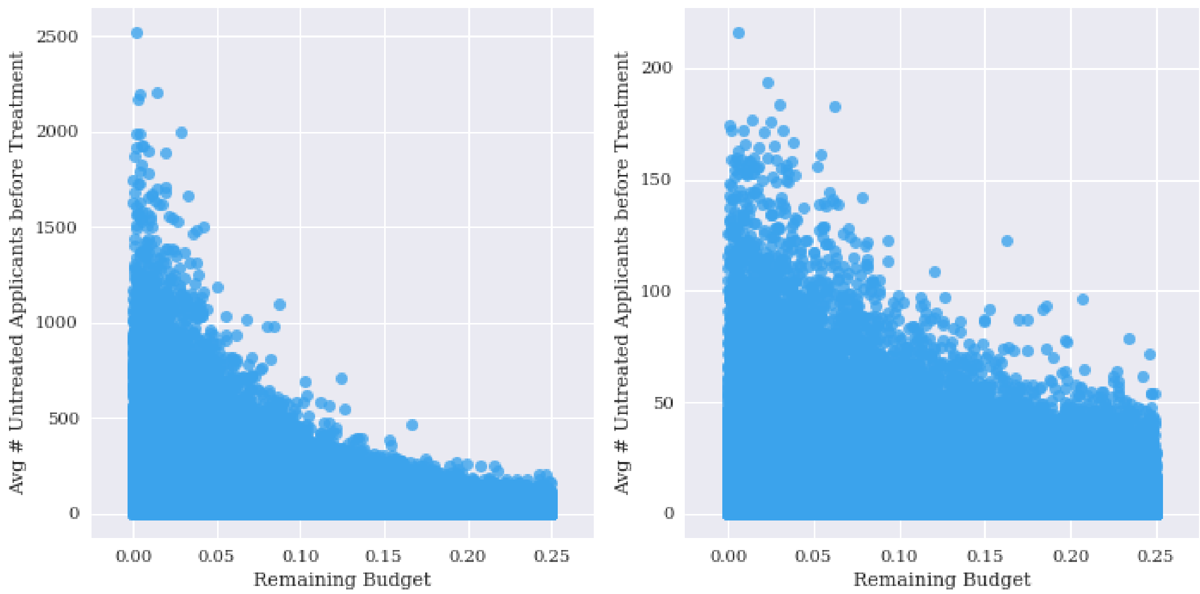


FIGURE G.2. Effect of Remaining Budget on Average Number of Rejected Individuals Prior to a Treatment (first and second order)



A: Doubly Robust Reward Estimates

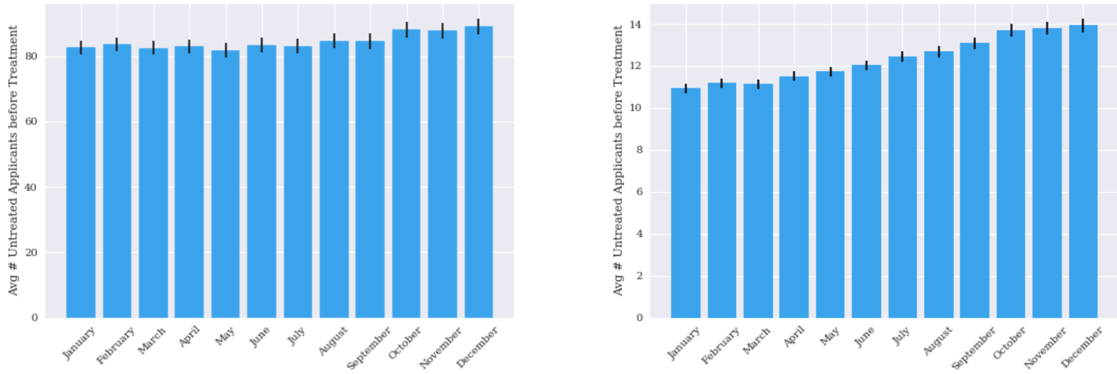
B: Standard OLS Reward Estimates

FIGURE G.1. Remaining Budget and Average Number of Rejected Individuals Prior to a Treatment across 100 Simulations

Table 2 shows again the correlation between the number of declined candidates with the remaining budget, but also with $cos(2\pi t)$ (i.e the effect of seasonality). For both cases, the correlation with $cos(2\pi t)$ is low: the depletion of budget appears to be the main driver of increased selectivity.

	Doubly Robust Rewards	Standard OLS Rewards
Remaining Budget	-0.492	-0.353
$\cos(2\pi t)$	0.007	-0.003

TABLE 2. Correlation of the Average Number of Rejected Individuals Prior to a Treatment with Time and Budget of the Policy Functions



A: Doubly Robust Reward Estimates

B: Standard OLS Reward Estimates

FIGURE G.3. Seasonal Differences in the Average Number of Rejected Individuals Prior to a Treatment

Figure G.3 provides further illustration of the policy function's behavior throughout the year. For the standard OLS rewards, an increasing number of rejections over the year is observable. As seasonality has mechanically been smoothed across seasons using the cosine function, the sharp difference between December and January in panel B is further evidence against the effect of seasonality. It does support the notion of increased selectivity with depleted budget as every episode starts on January 1st with a complete budget - i.e months later in the year are generally months with less budget. For the doubly robust rewards, this is less apparent - arguably due to the fact that episodes last for around 25 years and hence the different months are less associated with higher/lower budget.

The key difference between the optimal policy for doubly robust and standard OLS rewards appears to be the duration of a typical episode. We offer the following interpretation. While the two rewards concern the same dataset, they have entirely different distributions. A key difference is that there are more and more extreme outliers in case of the doubly robust rewards. In order to be relatively sure to avoid negative outliers, a higher selectivity is helpful. While this explains the longer episode-duration for the doubly robust rewards, it does not explain the large difference in the optimal policy with small versus large budget. We suspect that the latter is due to a high option value of budget for treating positive outliers.

Appendix H. An example with behavioral response to policy

In this section we present a simple example that illustrates how our techniques may be extended to a situation in which the policy affects the behavior of the individuals. Assume for simplicity that the setting is one of a finite budget, and that it is stationary in time (as in Section 2). Suppose further that there are two kinds of individuals, denoted by $l = 1, 2$. We shall separate l from the other covariates x. At any given point in time, the distribution of x in terms of the draw of arrivals is given by $F(x) := qF_1(x) + (1 - q)F_2(x)$, where q denotes the proportion of individuals from population 1, and F_1, F_2 denote the population distribution of the covariates in populations $1, 2$. Suppose now that

$$q = \Psi(\omega)$$

is actually a function of ω, where ω denotes the fraction of all the previously treated people who are from population 1. For instance, $\Psi(\cdot)$ could be strictly increasing, which means that the greater the fraction of population 1 in the treated population, the more people from population 1 are also likely to apply. Such a phenomenon could arise due to peer effects, for instance. Note that ω will now be a state variable even though it does not affect π_θ. Thus the state variables are $s := (x, z, \omega)$.

We will suppose that the policy maker is prohibited from discriminating on the basis of l, i.e π_θ is independent of l. Let

$$\begin{aligned} \bar{r}_{\theta,l}(z) &= E_{x\sim F_l}[r(x,1)\pi_\theta(1|x,z)], \quad \text{and} \\ \bar{G}_{\theta,l}(z) &= E_{x\sim F_l}[G_1(x,z)\pi_\theta(1|x,z) + G_0(x,z)\pi_\theta(0|x,z)], \end{aligned}$$

denote the population specific quantities. The net flow rates are thus given by

$$\begin{aligned} \bar{r}_\theta(z,\omega) &= \Psi(\omega)\bar{r}_{\theta,1}(z) + (1-\Psi(\omega))\bar{r}_{\theta,2}(z), \quad \text{and} \\ \bar{G}_\theta(z,\omega) &= \Psi(\omega)\bar{G}_{\theta,1}(z) + (1-\Psi(\omega))\bar{G}_{\theta,2}(z). \end{aligned}$$

Finally, we will also need a law of motion for ω. This is given by $\dot{\omega} = \Upsilon_a(s,l) := \mathbb{I}(l = 1 \cap a = 1)$. The expectation of $\Upsilon_a(s,l)$ conditional on z, ω will be denoted by

$$\bar{\Upsilon}_\theta(z,\omega) = \Psi(\omega)E_{x\sim F_1}[\pi_\theta(1|x,z)].$$

With the above definitions in mind, it is easy to see that we have the following PDE for the evolution of $h_\theta(z,\omega)$: (we take $\lambda = 1$ for simplicity)

$$\begin{aligned} \beta h_\theta(z,\omega) - \bar{G}_\theta(z,\omega)\partial_z h_\theta(z,\omega) - \bar{\Upsilon}_\theta(z,\omega)\partial_\omega h_\theta(z,\omega) - \bar{r}_\theta(z,\omega) &= 0 \text{ on } (0,z_0]\times[0,1], \\ h_\theta(z,\omega) &= 0 \text{ on } \{0\}\times[0,1]. \end{aligned} \tag{H.1}$$

This is nothing more than a PDE in two variables with a Dirichlet boundary condition.

For estimation, we can simply replace $F_1, F_2, r(x,1), \Psi$ with their estimated counterparts $F_{1n}, F_{2n}, \hat{r}(x,1), \hat{\Psi}$ and obtain a empirical version of PDE (H.1). Let $\hat{\theta}$ denote the resulting estimate after solving the empirical welfare maximization problem. Then as long as $\Psi(\cdot)$ is estimable at some rate n^{-c}, and $\bar{\Upsilon}_\theta$ is uniformly Lipschitz continuous, we can apply the techniques of this paper to show that

$$h_{\theta^*}(z,\omega) - h_{\hat{\theta}}(z,\omega) \lesssim \sqrt{\frac{v}{n}} + n^{-c}.$$

DEPARTMENT OF ECONOMICS, UNIVERSITY OF PENNSYLVANIA, USA, DEPARTMENT OF ECONOMICS, LONDON SCHOOL OF ECONOMICS, UK, DEPARTMENT OF ECONOMICS, UNIVERSITY OF ZURICH, SWITZERLAND.

E-mail address: akarun@sas.upenn.edu, f.c.geiecke@lse.ac.uk, claudio.schilter@econ.uzh.ch

Algorithm 5: Parallel Actor-Critic with non-compliance: Dirichlet boundary condition

Initialise policy parameter weights $\theta \leftarrow 0$
Initialise value function weights $\nu \leftarrow 0$
Batch size B

For $p = 1, 2, ...$ processes, launched in parallel, each using and updating the same global parameters θ and ν:

Repeat forever:

Reset budget: $z \leftarrow z_0$
Reset time: $t \leftarrow t_0$
$I \leftarrow 1$

While $(z, t) \in \mathcal{U}$**:**

$\theta_p \leftarrow \theta$ (Create local copy of θ for process p)

$\nu_p \leftarrow \nu$ (Create local copy of ν for process p)

batch_policy_upates $\leftarrow 0$

batch_value_upates $\leftarrow 0$

For $b = 1, 2, ..., B$**:**

$x \sim F_n$ (Draw new covariate at random from data)

hetero $\sim$ multinomial$(\hat{q_c}(x), \hat{q_a}(x), \hat{q_n}(x))$ (Draw compliance heterogeneity)

$a \sim \pi(a|s; \theta_p)$ (Draw action, note: $s = (x, z, t)$)

If hetero $= 1$ (Sample draw is a complier)

$R \leftarrow \widehat{LATE}(x) \cdot \mathbb{I}(a = 1)/b_n$ (I.e. $\hat{r}_n(s, a)$)

$z' \leftarrow z + G_a(x, z, t)/b_n$

Elseif hetero $= 2$ (Sample draw is always-taker)

$R \leftarrow 0$

$z' \leftarrow z + G_1(x, z, t)/b_n$

Elseif hetero $= 3$ (Sample draw is never-taker)

$R \leftarrow 0$

$z' \leftarrow z + G_0(x, z, t)/b_n$

$\omega \sim$ Exponential$(\lambda(t))$

$t' \leftarrow t + \omega/b_n$

$\delta \leftarrow R + \mathbb{I}\{(z', t') \in \mathcal{U}\} e^{-\beta(t'-t)} \nu_p^\intercal \phi_{z',t'} - \nu_p^\intercal \phi_{z,t}$ (TD error)

batch_policy_upates $\leftarrow$ batch_policy_upates $+ \alpha_\theta I \delta \nabla_\theta \ln \pi(a|s; \theta_p)$

batch_value_upates $\leftarrow$ batch_value_upates $+ \alpha_\nu I \delta \phi_{z,t}$

$z \leftarrow z'$

$t \leftarrow t'$

$I \leftarrow e^{-\beta(t'-t)} I$

If $(z, t) \notin \mathcal{U}$, break **For**

Globally update: $\nu \leftarrow \nu +$ batch_value_upates$/B$
Globally update: $\theta \leftarrow \theta +$ batch_policy_upates$/B$

Algorithm 6: Parallel Actor-Critic with clusters: Dirichlet boundary condition

Initialise policy parameter weights $\theta \leftarrow 0$
Initialise value function weights $\nu \leftarrow 0$
Batch size B
Clusters $c = 1, 2, \ldots, C$
Cluster specific arrival rates $\lambda_c(t)$

For $p = 1, 2, \ldots$ processes, launched in parallel, each using and updating the same global parameters θ and ν:

Repeat forever:

Reset budget: $z \leftarrow z_0$
Reset time: $t \leftarrow t_0$
$I \leftarrow 1$

While $(z, t) \in \mathcal{U}$**:**

$\theta_p \leftarrow \theta$ (Create local copy of θ for process p)

$\nu_p \leftarrow \nu$ (Create local copy of ν for process p)

batch_policy_upates $\leftarrow 0$

batch_value_upates $\leftarrow 0$

For $b = 1, 2, ..., B$**:**

$\lambda(t) \leftarrow \sum_c \lambda_c(t)$ (Calculate arrival rate for next individual)

$c \sim \text{multinomial}(p_1, \ldots, p_C)$ (where $p_c := \hat{\lambda}_c(t)/\hat{\lambda}(t)$)

$x \sim F_{n,c}$ (Draw new covariate at random from data cluster c)

$a \sim \pi(a|s; \theta_p)$ (Draw action, note: $s = (x, z, t)$)

$\omega \sim \text{Exponential}(\lambda(t))$

$t' \leftarrow t + \omega/b_n$

$z' \leftarrow z + G_a(x, z, t)/b_n$

$R \leftarrow \hat{r}(s, a)/b_n$ (with $R = 0$ if $a = 0$)

$\delta \leftarrow R + \mathbb{I}\{(z', t') \in \mathcal{U}\}e^{-\beta(t'-t)}\nu_p^\intercal \phi_{z',t'} - \nu_p^\intercal \phi_{z,t}$ (TD error)

batch_policy_upates $\leftarrow$ batch_policy_upates $+ \alpha_\theta I \delta \nabla_\theta \ln \pi(a|s; \theta_p)$

batch_value_upates $\leftarrow$ batch_value_upates $+ \alpha_\nu I \delta \phi_{z,t}$

$z \leftarrow z'$

$t \leftarrow t'$

$I \leftarrow e^{-\beta(t'-t)} I$

If $(z, t) \notin \mathcal{U}$, break **For**

Globally update: $\nu \leftarrow \nu +$ batch_value_upates$/B$
Globally update: $\theta \leftarrow \theta +$ batch_policy_upates$/B$