Using deep Q-learning to understand the tax evasion behavior of risk-averse firms

Nikolaos D. Goumagias[a,*], Dimitrios Hristu-Varsakelis[b], Yannis Assael[c],

[a]*Northumbria University, Newcastle Business School, Central Campus East 1, Newcastle upon Tyne, NE1 8ST, UK*
[b]*University of Macedonia, Department of Applied Informatics, Egnatia 156, Thessaloniki, 54006, Greece*
[c]*University of Oxford, Department of Computer Science, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK*

Abstract

Designing tax policies that are effective in curbing tax evasion and maximize state revenues requires a rigorous understanding of taxpayer behavior. This work explores the problem of determining the strategy a self-interested, risk-averse tax entity is expected to follow, as it "navigates" - in the context of a Markov Decision Process - a government-controlled tax environment that includes random audits, penalties and occasional tax amnesties. Although simplified versions of this problem have been previously explored, the mere assumption of risk-aversion (as opposed to risk-neutrality) raises the complexity of finding the optimal policy well beyond the reach of analytical techniques. Here, we obtain approximate solutions via a combination of Q-learning and recent advances in Deep Reinforcement Learning. By doing so, we i) determine the tax evasion behavior expected of the taxpayer entity, ii) calculate the degree of risk aversion of the "average" entity given empirical estimates of tax evasion, and iii) evaluate sample tax policies, in terms of expected revenues. Our model can be useful as a testbed for "in-vitro" testing of tax policies, while our results lead to various policy recommendations.

Keywords: Markov Decision Processes, Tax Evasion, Q-Learning, Deep Learning

1. Introduction

While the aftershocks of the latest global financial crisis are still being felt, many governments struggle to implement public policy because of budget deficits or lagging tax revenues (Bayer et al., 2015). The latter problem arises as a result of reduced economic activity, or when there is a strong sense among taxpayers that the expected personal benefit from tax evasion surpasses the corresponding social benefit of paying taxes (Alm and Beck, 1990; Bornstein and Rosenhead, 1990). This, in the absence of a properly designed tax system and enforcement mechanism, leads to tax evasion, a serious crime that saps the State of revenue and undermines the sense of social justice, as dishonest taxpayers seem to enjoy the same public goods as honest ones do. The resulting "shadow economy" also has a strong adverse impact on credit ratings and lending costs (Markellos et al., 2016), welfare programs, fiscal policies and unemployment (Fleming et al., 2000).

Of course, tax systems typically contain various safeguards to discourage tax evasion (defined here as the deliberate failure to declare all or some of ones income to tax authorities). In practice, however, tax systems are rather complex policy structures which are difficult to make "air-tight" in terms of tax evasion, for reasons having to do with i) occasional ambiguity in tax regulations, which hinders tax compliance and enforcement (Andreoni et al., 1998), and ii) the heterogeneous behaviors of the various taxpayer entities, based on their individual risk preferences (Hokamp and Pickhardt, 2010).

This paper is concerned with the development of a rigorous computational framework which can describe and predict the behavior of tax evaders, assuming that they are self-interested and work to maximize the utility of their own revenues, balancing potential gains from tax evasion against the risk of getting caught. In particular, we are interested in i) estimating State revenues for any given set of tax parameters (e.g., tax rates and penalties), ii) testing whether specific tax regulations are helpful or not, and iii) predicting how taxpayers - and tax compliance - respond to parameter changes. This last item is linked to taxpayers' risk aversion, knowledge of which would help the State determine the effects of, for example, an increase in tax penalties or audit rates.

The issues raised above are essential to the State if it

*Corresponding author
Email addresses:
nikolaos.goumagias@northumbria.ac.uk (Nikolaos D. Goumagias), dcv@uom.gr (Dimitrios Hristu-Varsakelis), yannis.assael@cs.ox.ac.uk (Yannis Assael)

is to know the extent to which its tax policies are working or to rank alternative policies and take steps towards maximizing revenue. In this work, we propose to explore them using a combination of deep neural networks and Q-learning for determining the tax evasion behavior of a risk averse taxpayer (we will use the term "firm" henceforth because we will be interested mainly in business entities). We will develop and test our approach in a context that builds on the work of Goumagias et al. (2012) (where only the case of risk neutrality was analyzed) and involves a close-to-real-world tax system, with many of the usual trappings such as tax rates, random audits, penalties and occasional tax amnesties, as well as taxpayer heterogeneity. As we shall see, the introduction of risk aversion into the model and the resulting nonlinearity of the firm's utility function combines with the firm's dynamics and leads to a significant increase in complexity. This puts the problem finding the firm's optimal behavior well beyond the reach of analytical methods and requires powerful approximation techniques to be brought to bear.

The main contributions of this work are i) the use of the deep reinforcement learning techniques to obtain computational solutions for the firm's optimal behavior based on the Markov dynamical model of Goumagias et al. (2012) and ii) a computational framework for exploring the behavior expected of self-interested risk-averse firms who may choose to engage in tax evasion in order to maximize their own utility. In addition, and on more practical grounds, we estimate the risk aversion coefficient of the "average" firm - or group of firms - given empirical data on its tax compliance and evaluate sample tax policies in terms of their benefit for the State (or, equivalently, the level of tax evasion they result in). To our knowledge, ours is the first work to apply deep learning in the context of taxation and tax evasion, and the first to obtain solutions that reveal the behavior of a risk-averse firm at a "fine" timescale, i.e., on a year-to-year basis, based on its evolving status in the "eyes" of tax authorities. We view our approach as particularly relevant both in light of the growing interest in deep learning applications and for the opportunities that our model affords to regulators in the design of effective policies that make entities behave more honestly.

The remainder of this paper is structured as follows. In Section 2 we review the relevant literature and discuss how our approach is situated relative to previous work. Section 3 begins with a brief description of the tax system in which the firm operates, and explains the main parameters. In the same Section we describe a Markov-based model of the firm's evolution through the tax system and pose the main optimization problem we are interested in solving and the computational challenges involved. Our solution approach, combining Q-learning and Deep Neural Networks, is detailed in Section 4. Finally, Section 5 discusses the results we obtained - using the Greek tax system as a case study for the sake of concreteness - and their relevance to the questions posed above regarding the firm's expected behavior, incentives for reporting profits, degree of risk aversion, and policy implications.

2. Related work

Prior work related to optimal taxation and tax-evasion modeling can be grouped into two main categories: i) analytic (macroeconomic, and principle-agent based), and ii) computational (agent-based, simulation-based). The seminal work in the first category was Allingham and Sandmo (1972) who introduced a model of optimal taxation posed as a portfolio allocation problem. Several scholars built on that model by also introducing labor supply (Yitzhaki, 1974; Baldry, 1979) and public goods offered (Cowell, 1981). The complexity of the phenomenon was highlighted early on by Clotfelter (1983) and Crane and Nourzad (1986), who challenged the monotonic relationship between tax rates and tax evasion. One of the drawbacks of the analytical approaches was that they often implied less behavioral heterogeneity on behalf of taxpayers than what was suggested by empirical evidence (Andreoni et al., 1998), and - in order to remain tractable - they could not fully capture the dynamics of tax evasion (Martinez-Vazquez and Rider, 2005).

In particular, beyond the issue of accounting for heterogeneity (e.g., in taxpayers' risk-aversion), there exists much interesting structure in taxpayers' behavior if one considers "fine-grained" models of their evolution through the tax system. In that setting, one must reckon with the various random transitions the taxpayer may undergo from year to year, such as being audited, or offered the chance to participate in a tax amnesty program (we will provide details on such options shortly), or changing preferences via interaction with others. Such considerations have led to a number of recent computational-based approaches in the form of automaton-based (Garrido and Mittone, 2012) and agent-based (Gao and Xu, 2009) models. Computational approaches may allow for more realism, by having, for example, a large number of agents interact with each other based on predetermined characteristics related to the taxation parameters and intrinsic utility functions (Pickhardt and Seibold, 2014). Their advantage is that they can offer empirically grounded and

theoretically-informed policy implications, but they often suffer from a limited analytical tractability of the solutions they suggest.

An attempt to overcome these limitations while modeling the year-to-year behavior of the firm was made by Goumagias et al. (2012), who introduced a parametric Markov-based model describing the evolution of a rational firm within the Greek tax system. The firm's goal was to maximize a discounted sum of its yearly after-tax revenues, possibly by engaging in tax evasion. That work showed that the firm would attempt to evade taxation as much as possible under the system currently in place, and produced "maps" showing which combinations of tax parameters lead firms to behave honestly and which do not. A severe limitation of Goumagias et al. (2012) was the fact that it applied only to the special case of *risk neutral* entities. That assumption kept the firm's state and decision spaces conveniently small (it implied, for example, that the firm's optimal decision is to either be completely honest or to conceal as much profit as possible, eliminating "intermediate" options), making the problem of optimizing taxpayer behavior solvable via Dynamic Programming (DP) (Bertsekas, 1995). Of course, most taxpayer entities are not likely to be risk neutral; thus, it becomes necessary to incorporate risk-aversion into the analysis in order to be able to predict the behavior of a broad spectrum of taxpayers and explore the effectiveness of tax policies in a more realistic setting.

As we will discuss in Sec. 3.3, risk-aversion introduces nonlinearity in the firm's objective function, making analytical or DP methods ineffective, and we will require some way of circumventing the curse of dimensionality in that context. Among the various alternatives, iterative dynamic programming can potentially allow for tractable solutions (Jaakkola et al., 1994), however, that method's applicability is limited when faced with multiple sources of uncertainty, as is the case here. Computational solutions, including artificial intelligence methods and neural networks for cost-to-go function approximation (Tsitsiklis and Van Roy, 1996; Wheeler and Narendra, 1986; Watkins, 1989) will prove to be more promising in our setting. Reinforcement learning-based methods, in particular, approximate the cost-to-go function via simulation and perform function approximation via regression or neural networks (Gosavi, 2004). This approach includes algorithms such as R-learning (Singh, 1994; Tadepalli and Ok, 1996), and Q-Learning (Sutton and Barto, 1998; Tsitsiklis, 1994). One advantage of reinforcement learning which will be useful to us is that, unlike DP, the process can be set to update the value of the cost-to-go function for the states that are most often visited (Tsitsiklis, 1994).

Recently-proposed *deep learning* algorithms have greatly broadened the scope of applicability of artificial intelligence and machine learning, beyond "classical" problems of pattern recognition (LeCun et al., 2015) and have shown great promise in approximating complex nonlinear cost-to-go functions (Schmidhuber, 2015). To date, deep learning has been applied to challenging problems in areas including image recognition and processing (Krizhevsky et al., 2012), speech recognition (Mikolov et al., 2011), biology (Leung et al., 2014), analysis of financial trading (Krauss et al., 2017), social networks (Perozzi et al., 2014) and human behavior (Ronao and Cho, 2016). Here, we will make use of recent developments in deep reinforcement learning in order to obtain computational solutions for the firm's optimal behavior, with all of our model's complexities. This opens the door to more informed policy decisions by providing a computational platform for comparing tax policies (e.g., those with tax amnesty vs those without), estimating the firms' degree of risk aversion from empirical data, predicting the expected tax revenue for the government, or calculating the effects of a change in any tax parameter on revenues.

3. Model description

We proceed with a brief discussion of the tax system within which the firm evolves, to be followed by the corresponding mathematical model. That model will be parametric, with many of the tax "features" commonly encountered, including random audits and penalties. Of course, when it comes time to make computations, we will have to select parameter values (tax rates, etc.) for a specific locale. We will focus on Greece in particular, for the sake of concreteness and because, with tax evasion being a significant and long-standing problem there, one can draw interesting and practical conclusions. However, the basic tax provisions we consider appear in most tax systems, and our model could be adjusted to describe matters in other countries as well.

3.1. A basic taxation system with occasional optional amnesties

The basic components of our taxation system will include - as is the case in most countries - a tax rate on profits, random audits for identifying tax evaders, and monetary penalties for under-reporting income. Those penalties, added to the original tax due on any unreported income discovered during an audit, will be proportional to the amount of unreported income and the

time elapsed since the offense took place. We will also allow any penalty to be discounted somewhat for prompt payment. The tax authority will audit a small fraction of cases each year but will retain the right to audit a firm's tax returns for a number of years in the past. Any tax-evasion activity beyond that horizon will be considered to be beyond the statute of limitations.

Our model will also include an optional tax amnesty in which the government may occasionally allow taxpayer entities to pay a fee in exchange for which past tax declarations are closed to any audits. This "closure" fee will be paid separately for each tax year a firm would like to exempt from a possible audit. It is worth noting that the appeal of tax amnesties as revenue collecting mechanisms is typically reinforced during and after long recessions (Ross and Buckwalter, 2013; Bayer et al., 2015). Amnesties are more commonly used than one might expect. For instance, only in the US, between 1982 and 2011, there were 104 cases of some form of tax amnesty (Ross and Buckwalter, 2013). Other examples include India (Das-Gupta and Mookherjee, 1995), and Russia (Alm and Rath, 1998). In Greece, the closure option mentioned above was being offered roughly every 4-5 years during 1998-2006 (e.g., Hellenic Ministry of Finance (2004) and Hellenic Ministry of Finance (2008)). More recently, it was re-introduced in the Greek parliament with a new round being under consideration (Hellenic Ministry of Finance, 2015). The irregular usage of tax amnesties as tax revenue collection mechanisms increases the complexity of decision making both on behalf of the government and the taxpayer. The use of tax amnesties by firms essentially shrinks the audit pool. Thus, if in some year the government offers the closure option but a firm refuses to use it, that firm is more likely to be audited. For a more detailed explanation of the mechanics of closure, see Goumagias et al. (2012). In practical terms, one question we would like to answer is whether such a measure (although it provides some immediate tax revenue) actually hurts long-term revenues because it might act as a counter-incentive to paying the proper tax (Bayer et al., 2015).

3.2. The behavior of risk-averse firms with optional closure

The work in Goumagias et al. (2012) codified the firm's time evolution through the tax system described above, in a compact Markov-based model which includes all of the basic features described in Sec. 3.1, including tax rates, penalties, a five-year statute of limitations for audits of past tax statements, and occasional tax amnesty (closure). We will revisit it here briefly, in as compact form as possible, and extend it for our purposes.

For a tax system with a five-year statute of limitations on auditing past tax statements, the firm's evolution can be described by the linear state equation (Goumagias et al., 2012)

$$x_{k+1} = Ax_k + Bu_k + n_k, \tag{1}$$

where $x(0)$ is given, and $A \in \mathbb{R}^{7\times7}$, $B \in \mathbb{R}^{7\times2}$, $n_k \in \mathbb{R}^7$ are as in Appendix A.

The firm's state at discrete time k is given by the triple $x_k = [s_k, c_k, h_k^T]^T \in \mathcal{S} \times \{1, 2\}, \times[0, 1]^5$. Here, $\mathcal{S}$ is a 15-element set (in the discussion that follows, it will be convenient to use $\mathcal{S} = \{1, ..., 15\}$), containing the firm's possible tax statuses (see Goumagias et al. (2012) for a graphical explanation): the first five elements of $\mathcal{S}$ correspond to the firm currently being audited, with 1-5 years since its last audit (any tax declarations "older" than 5 years are beyond the statute of limitations); elements 6-10 correspond to the firm using the closure option with 1-5 years having passed since its last audit or closure; and states 11-15 correspond to the firm being unaudited for 1-5 years (not being currently audited, nor using closure). Of the remaining state elements, c_k is a two-level variable denoting whether the government has made the closure option available at time k, and h_k contains the time history of the firm's past 5 decisions with respect to tax evasion, with elements in h ranging from 0 (full disclosure) to 1 (the firm hides as much of its income as possible).

In Eq. 1, u_k is a 2-element vector containing the firm's actions in year k; the first element, $[u_k]_1 \in [0, 1]$ denotes the fraction of profits that the firm decides to conceal, while the second, $[u_k]_2 \in \{0, 1\}$ is a binary decision on whether or not to use the closure option, if it is available. In the term $n_k = [\omega_k, \epsilon_k, 0_{5\times1}]^T$, ω_k determines the first element of the "next" state vector, i.e., the firm's status in the tax system (e.g., being audited or not, or removing itself from this year's audit pool by making use of the closure option), according to a Markov decision process whose transition probabilities depend on the current state and the firm's decision to use closure (see Goumagias et al. (2012), also given in Appendix B to facilitate review). The ϵ_k are Bernoulli-like, taking on the value 2 when the government offers the closure option (this is assumed to occur with some probability p_0), or 1 otherwise.

The firm "weighs" its rewards (profit, plus any taxes it is able to save by declaring less of it) according to a constant relative risk aversion utility function

$$U(z) = \frac{z^{1-\lambda}}{1-\lambda}, \tag{2}$$

with λ being the associated risk-aversion coefficient, and $z = g(x_k, u_k)$ being the reward the firm receives when in state x_k and taking an action u_k. Based on the earlier description of the rules of the tax system, $g(\cdot,\cdot)$ is given by

$$
\begin{aligned}
g\left(x_k,\, u_k\right) &= g\left(\left[s_k,\, c_k,\, h_k^T\right]^T,\, u_k\right) \\
&= R \cdot \begin{cases} \left(1 - r + r\left[u_k\right]_1\right), & s_k \in \{11-15\} \\ \left(1 - r + r[u_k]_1 - \ell\left(s_k - 5\right)\right), & s_k \in \{6-10\} \\ \begin{pmatrix} 1 - r + r\left[u_k\right]_1 \\ -r\sum_{i=1}^{s_k}\left[h_k\right]_{6-i} \end{pmatrix}, & s_k \in \{1-5\} \\ \begin{pmatrix} -r\beta_d\beta\sum_{i=1}^{s_k} i\left[h_k\right]_{6-i} \end{pmatrix} \end{cases}
\end{aligned} \tag{3}
$$

where R denotes the firm's annual revenues, r is the tax-rate, ℓ the closure cost (paid if the firm decides to take advantage of that option in the event it is offered), β the tax-penalty and β_d is the discount factor for prompt payment. In Eq. 3, the top term corresponds to the firm's reward if it is not audited, so that depending on $[u_k]_1$, it may pay all to none of the tax due. In the middle term, the firm is using the closure option, so that it pays ℓ for as many years as it has gone unaudited, up to a maximum of five. Finally, the bottom term in Eq. 3 corresponds to the firm being audited, so that it pays any back taxes due (based on its historical behavior) and the corresponding penalties, as per our earlier description.

The firm is assumed to act in a self-interested way and thus chooses its policy u_k so as to maximize the discounted expected reward:

$$
\max_{u_k} \mathcal{E}_{\omega_k,\, \epsilon_k}\left\{\sum_{k=0}^{\infty}\gamma^k U\left(g(x_k, u_k)\right)\right\} \tag{4}
$$

where $\gamma \in (0, 1)$ denotes the discount factor.

It can be shown (in a way similar to Goumagias et al. (2012)) that the Bellman equation whose solution maximizes (4) is equivalent to

$$
\begin{aligned}
J_\infty\left(i, q, h\right) = \max_u \Big\{ & U\left(g\left(i, q, h, u\right)\right) \\
& + \gamma \sum_{t=1}^{2}\sum_{j=1}^{15} P_{\text{qji}}\left([u]_2\right) \cdot Pr(\epsilon = t) \cdot \\
& \cdot J_\infty(j, t, Hh + [0\,0\,0\,0\,1]^T\,[u]_1)\Big\}
\end{aligned} \tag{5}
$$

where, for convenience, we have slightly abused the notation by writing $J_\infty(i, q, h)$ instead of $J_\infty(x)$, with $i \in S = \{1, \ldots, 15\}$, $q \in \{1, 2\}$, and $h \in [0, 1]^5$.

3.3. Challenges in solving for the firm's expected strategy

There is a significant difficulty when it comes to solving Eq. 5 for the optimal firm reward (and the associated tax-evasion policy), stemming from the continuity of certain elements in the state and control vectors. As we have already mentioned, the first element, $[u_k]_1 \in [0, 1]$, of the control vector u_k denotes tax-evasion as fraction of the firm's annual revenues. This implies that u as well as x are continuous because the firm's last five tax-evasion decisions are always incorporated into the state. This makes Eq. 5 difficult to compute.

One may attempt to circumvent this problem by discretizing the variables in question to render both the state and the control vector discrete. For example, we may instead consider $[u_k]_1 \in [0, 0.01, \ldots, 0.99, 1]$, and assume that tax-evasion takes place in increments of 1%, which seems like a reasonable level of coarseness. However, after thus discretizing the control and state spaces, the number of state-control pairs, (x, u), remains large. Specifically, we are left with $15 \times 2 \times 101^5 \times 202$ potential pairs (the number of the elements of the state vector x_k including all possible combinations of control for the past five years, times the number of possible controls in u_k). Such a number of states is too large for DP to be effective in solving the stationary Bellman equation via value iteration, for example, because: i) "visiting" every state in order to update the value function associated with Eq. 5 becomes infeasible and ii) it is difficult to even store the function $J(x, u)$ (the value of applying decision u while at state x, as a precursor to computing the maximum in the above equation) in tabular form, as one would have to do if Eq. 5 were to be solved via value iteration, for example.

The work in Goumagias et al. (2012) circumvented these difficulties by assuming risk-neutrality ($\lambda = 0$) on behalf of the firm (and thus linearity of the reward function) and successfully applied DP after determining that $[u_k]_1$ should only take a "bang-bang" form (conceal as much revenue as possible or none at all), leading to a significant reduction in the number of state-control pairs. In our case, however, the cost-*to-go* function (Eq. 3) is non-linear, so that we must consider the full range of control values, and it is thus computationally difficult to apply DP.

One way to go forward is to combine: i) an approximation method to estimate the value function J_k and ii) an approximate way of storing the optimal values of J_k, based on the optimal policy. To address the former we will use reinforcement learning – specifically Q-learning, as described in Sutton and Barto (1998), where

J_k will play the role of the Q-function $Q(x_k, u_k)$, while for the latter, a deep Artificial Neural Network will be used, as we will discuss shortly.

4. Constructing an approximator: Deep Q-Learning

We experimented with various choices of learning algorithms and neural network architectures for the purposes of learning and storing the optimal value function given in the previous Section. In the following we describe our solution, combining Q-learning and a Deep Neural Network, and discuss some of the difficulties involved and how they can be overcome.

4.1. Q-learning

Q-learning is a model-free reinforcement learning method (Sutton and Barto, 1998), that is used to find an optimal action-selection policy for any given finite MDP. In the "language" of Sutton and Barto (1998), an agent (in our case the firm) observes the current state $x_k \in \mathcal{X} = \mathcal{S} \times C \times [0,1]^5$ at each discrete time step k, chooses an action $u_k \in \mathcal{U} = [0,1] \times \{0,1\}$ according to a possibly stochastic policy π, mapping states to actions, observes the reward signal $U(g(x_k, u_k)) \in \mathbb{R}$, and transitions to a new state x_{k+1}. The objective is to maximize an expectation over the discounted return, as in Eq. 4.

Briefly, Q-learning involves sequentially updating an approximation of the action-value function, i.e., the function that produces the expected utility of taking a given action at a given state and following the optimal policy thereafter. The so-called Q-function of a policy π is $Q^\pi(x, u) = \mathcal{E}\{D_k | x_k = x, u_k = u\}$, where

$$D_k = \sum_{i=0}^{\infty} \gamma^i U(g(x_{k+i}, u_{k+i})), \tag{6}$$

and the state evolution proceeds under the policy π. Finally, the optimal action-value function $Q^*(x, u) = \max_\pi Q^\pi(x, u)$ to which the learning process is to converge, obeys the Bellman Eq. 4.

For our purposes, in the notation of Sec. 3, the function J we are seeking (5) is simply the Q^* function, after having maximized over u. Common choices for modeling the Q-function are lookup tables and linear approximators, among others. However, these models suffer from poor performance and scalability problems, and cannot possibly handle the high-dimensional state space involved in our case, as we discussed in Sec. 3.3. An efficient alternative to the aforementioned models are neural networks.

4.2. Deep Q-Networks (DQN)

Deep Q-learning (DQN) was introduced by Mnih et al. (2015), and uses neural networks parametrized by θ to represent $Q(x, u; \theta)$, where the Q function is augmented with a parameter vector θ, usually consisting of the weights and biases of the multiple layers of the network. Neural networks, viewed as general function approximators, are trained "end-to-end", and can efficiently handle high-dimensionality problems. Recently, a DQN surpassed human performance in 49 different Atari games (Mnih et al., 2015). For our purposes, the DQN will receive as input the firm's state x_k and will have to produce the optimal decision, u_k. Because the network will be trained to capture the optimal firm policy, we will sometimes refer to it as the "policy network".

DQNs are trained iteratively using stochastic gradient descend, until convergence. This is done by minimizing, at each iteration i, a loss function of the network's parameters, $\mathcal{L}_i$, which is expressed as

$$\mathcal{L}_i(\theta_i) = \mathcal{E}_{x,u,r,x'}\left\{\Delta Q^2\right\}, \text{with} \tag{7}$$

$$\Delta Q = Y^{DQN} - Q(x, u; \theta_i), \text{and} \tag{8}$$

$$Y^{DQN} = U(g(x_k, u_k)) + \gamma \max_{u'} Q(x', u'; \theta_i^-), \tag{9}$$

and θ_i^- is an "older" copy of the network's parameters, as we explain next. Function approximation using neural networks can be unstable, and we observed such behavior in our numerical experiments, particularly after we introduce a second source of uncertainty in the form of closure availability. Following Mnih et al. (2015), to stabilize the process we use a so-called "target network", i.e., a copy of our original DQN which has the same architecture but a different set of parameters, θ_i^-. The parameters of the target network represent an older version of the policy network and are updated at a slower rate. Thus, while the policy network acts to produce inputs u that will steer the firm to its next state, the slowly-updated target network is used to compute Y^{DQN} which, in turn, is used to improve the parameters of the policy network via gradient descent:

$$\nabla_{\theta_i} \mathcal{L}_i(\theta_i) = \mathcal{E}_{x,u,r,x'}\left\{\Delta Q \nabla_{\theta_i} Q(x, u; \theta_i)\right\}. \tag{10}$$

While training the DQN, we must choose an action u to drive the state at each iteration. That action is to be chosen from $Q(x, u; \theta_i)$ using an ϵ-greedy policy that selects the u that maximizes Q with probability $1 - \epsilon$, or a random u with probability ϵ. Additionally, our DQN uses so-called "experience replay" (Lin, 1993). During learning, we maintain a set of episodic experiences (tuples that include the state, the action taken, the resulting

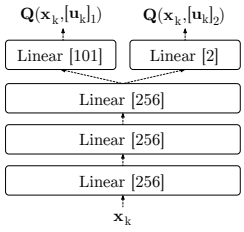


Figure 1: Schematic representation of our DQN. We use a 4-layer network that takes as input the current state x_k. The first three layers consist of 256 neurons, followed by two parallel layers of 101 and 2 neurons, for computing $Q(x_k, [u_k]_1)$ and $Q(x_k, [u_k]_2)$, respectively.

state and reward received). The DQN is then trained by sampling mini-batches of those experiences. This has the effect of stabilizing the learning process and avoids overfitting. Experience replay was used very successfully by Mnih et al. (2015) and it is often motivated as a technique for reducing sample correlation, while also enabling re-use of past experiences for learning. Furthermore, it is a valuable tool for improving sample efficiency and can also improve performance by a significant margin, as it did in our case.

A final but important modification was the use of Double Q-learning, a technique introduced very recently by van Hasselt et al. (2016a). Double Q-learning for DQN (DDQN) reduces overestimation of the Q-values by decomposing the max operation in the target network into action selection and action evaluation. Thus, instead of using the target network's maximum Q-value estimate in Eq. 9, we use the target network's Q-value of the current network's best action. The DDQN update equations are the same as for DQN, after replacing the target Y^{DQN} in Eq. 9 with

$$Y^{DDQN} = U\left(g\left(x_k, u_k\right)\right) + \gamma Q(x', \arg\max_{u'} Q(x', u'; \theta_i); \theta_i^-). \tag{11}$$

The entire Double DQN training loop is given in pseudocode in Algorithm 1 below.

4.3. DQN architecture

Our network architecture was inspired by the model of Mnih et al. (2015). The action-space described in Sec. 3 consists of two action elements $[u]_1$ and $[u]_2$. The firm's tax-evasion level is determined by $[u]_1 \in [0, 1]$, discretized in intervals of 1% resulting a set of 101 actions. This convention is commonly used to take advantage of the off-policy stability of Q-learning compared to on-policy $SARSA - \lambda$, actor-critic or policy gradient approaches. The firm's use of the closure option is $[u]_2 \in \{0, 1\}$, and if closure is not available then $[u]_2 = 0$.

Algorithm 1: Double DQN (van Hasselt et al., 2016a)

▷ Initialize experience replay memory D,
action-value function Q with random weights θ
and set $\theta^- = \theta$.
for *episode = 1 to M* **do**
  for *k = 1 to K* **do**
    `// Take an action`
    Select u_k randomly with probability ϵ else
    $\arg\max_u Q(x_k, u; \theta)$
    Execute u_k and observe reward
    $U\left(g\left(x_k, u_k\right)\right)$ and state x_{t+1}
    Store transition $\left(x_k, u_k, U\left(g\left(x_k, u_k\right)\right), x_{t+1}\right)$
    in D
    `// Training step`
    Sample minibatch (x_j, u_j, r_j, x_{j+1}) from D
    $Y^{DDQN} = U\left(g\left(x_j, u_j\right)\right) +$
    $\gamma Q(x', \arg\max_{u'} Q(x', u'; \theta_i); \theta_i^-)$
    Perform a gradient descent step
    $\nabla_\theta (Y^{DDQN} - Q(x_j, u_j; \theta))^2$
    `// Update target network`
    Every C steps reset target network, i.e., set
    $\theta^- = \theta$
  end
end

Our approximator (see Fig. 1) is a 4-layer multilayer perceptron (MLP) and takes as input the current state x_k. The first three layers consist of 256 neurons, followed by two parallel linear layers of 101 and 2 neurons, for computing $Q(x_k, [u_k]_1)$ and $Q(x_k, [u_k]_2)$, respectively. The network makes use of the rectified linear unit (ReLU) transformation function $f(x) = \max(0, x)$ between layers.

Finally, our setting requires the DQN to produce two action elements $([u]_1, [u]_2)$. To improve the scalability of our approximator, and after numerical experimentation, we opted to use independent Q-learning to learn two different Q-functions (one for each component of the firm's decision, $[u]_1$ and $[u]_2$, as in (Narasimhan et al., 2015; Foerster et al., 2016). In this case, the DQN loss is expressed as

$$\begin{aligned}
\Delta Q_{[u]_1} =& U\left(g\left(x_k, u_k\right)\right) \\
&+ \gamma max_{[u_{k+1}]_1} Q(x_{k+1}, [u_{k+1}]_1) - Q(x_k, [u_k]_1) \\
\Delta Q_{[u]_2} =& U\left(g\left(x_k, u_k\right)\right) \\
&+ \gamma max_{[u_{k+1}]_2} Q(x_{k+1}, [u_{k+1}]_2) - Q(x_k, [u_k]_2) \\
\mathcal{L}_i(\theta_i) =& \mathcal{E}_{x,u,r,x'}\left\{\Delta Q_{[u]_1}^2 + \Delta Q_{[u]_2}^2\right\}.
\end{aligned} \tag{12}$$

5. Evaluating the model: results and discussion

As we have mentioned in the Introduction, we are generally interested in being able to evaluate the firm's decisions (assuming that it acts in a self-interested way) - and maximum expected utility under various degrees of risk-aversion, thereby producing a tool that could be used to predict firm behavior, compute tax revenue, and to gauge the reaction of the firm to tax policy scenarios under consideration by the government. We are also interested in characterizing the firm's strategy by determining, for example, whether the firm would be expected to use a constant degree of tax-evasion ($[u]_1$) in every state (as in Goumagias et al. (2012)), finding the firm's coefficient of risk-aversion given empirical estimates of the degree of tax evasion, and examining whether it is beneficial for the government to offer the closure option in any of the settings discussed in the Introduction.

5.1. Model parameters and Training setup

The various tax parameters present in our model were selected using Greece as a case study for the sake of concreteness, to facilitate comparisons with prior work (Goumagias et al., 2012), and because that country presents an interesting case as it is plagued by widespread tax evasion (we will discuss estimates in Sec. 5.4.1). Specifically, the tax and audit rates were $r = 0.24$ and $p = 0.05$, respectively; the statute of limitations for auditing past tax statements was 5 years; the penalty for underreported profit was $\beta = 0.24$ (24% annually); potential tax penalties were discounted by 40% if paid immediately ($\beta_d = 0.6$); and, finally, the cost for the firm to use the closure option - if available - was $\ell = 0.023$.

Training our DQN-based model to optimize the firm's behavior for any one set of parameters (risk-aversion coefficient, closure probability and cost, audit probability, penalty coefficient) required about 2 days on an Intel® Xeon® X5690 CPU with 72GB of RAM. Our source code is freely available under an open-source license at `https://github.com/iassael/tax-evasion-dqn`. The network was trained on $50,000$ episodes of the firm's evolution, each lasting 250 time steps. The network's performance was evaluated every 100 episodes as the average discounted reward of those episodes. We followed the training methodology proposed by Mnih et al. (2015), using Double Q-Learning (van Hasselt et al., 2016a). Because $x_k \in [0, 1]^{21}$, the inputs to the network were "shifted" by subtracting 0.5 from all elements of the state x_k. Shifting the inputs to be evenly spread around 0 resulted in faster convergence[1].

As usual, the network's training objective was to minimize the mean squared temporal difference error. Thus, the backpropagated gradients described above were significantly affected by the scale of the rewards. Looking at the form of the risk-averse utility function $U(\cdot)$ in Eq. 2, this becomes problematic for input values close to 0, where U dives to $-\infty$. To stabilize the training process numerically, the values returned by U were clipped below, so that they always lie in $[-1, 0)$. That is, if the argument of U was less than ϵ_{thresh}, where $U(\epsilon_{thresh}) = -1$, the argument was replaced by ϵ_{thresh}. Our empirical evaluation showed that reward clipping was crucial to deal with the steep non-linear scale of rewards. The particular value of -1 was not critical - more negative values work just as well, as long as they are "far" from the utility values the firm usually operates around, but not too negative so as to end up in extremely steep parts of U near zero.

Our ϵ-greedy exploration policy used $\epsilon = 0.5$ which linearly decreased to $\epsilon = 0.1$ in the first 5000 episodes. This resulted in a highly-explorative policy in the beginning which rapidly converged to a more exploitative one. The training process took advantage of past experiences, as we explained above (experience replay with mini-batches of size 100), and the target network described in Sec. 4.1 was updated every 10 episodes. The networks' parameters were optimized using Adam (Kingma and Ba, 2014) with a learning rate of 10^{-4}.

We proceed by first evaluating our model in the case of a risk neutral firm - for the purposes of comparison with prior work. Following that, we will discuss the case of a risk-averse firm and will explore its behavior.

5.2. Risk-neutral firms: comparison with known optimum.

Before attempting to compute a risk-averse firm's expected behavior, we validated our approach against the known optimal solution for risk-neutral firms from Goumagias et al. (2012). Tab. 1 shows the firm's total discounted rewards in four cases which are of interest, according to how often the closure option is offered each year: a) never, b) with probability 0.2, c) always, and d) periodically, every 5 years.

[1] A simple example where this type of shifting improves learning is the case of one-hot encoded inputs x, where both the weights W, and biases b, of the network can be being "learned" even when the original inputs are zero, i.e., $f(x) = ReLU(Wx + b)$, whereas without shifting, only b would be learned when $x = 0$.

Closure Option	Dynamic Programing	DQN
Never	3254.6	3270.66
$p_{closure} = 0.2$	3307.9	3316.76
Always	3358.3	3357.01
5-periodic	3319.7	3335.75

Table 1: Total discounted revenue for risk-neutral firm, as computed by our model vs. via Dynamic Programming as reported in Goumagias et al. (2012).

Our DQN approach is inherently an approximate one. We note however that the firm revenues we computed differ less than 0.5% from the "true" values computed via DP. Besides the optimal firm revenues, the optimal firm policies were identical to those found in Goumagias et al. (2012) in each of the four cases examined, i.e., it was always optimal for the firm to conceal as much profit as possible and to make use of the option whenever available.

5.3. The behavior of risk-averse firms - ranking sample tax policies

We performed a series of runs designed to explore the effect of risk aversion on the behavior of the firm, by keeping the tax-parameters fixed to the values mentioned in Sec. 5.1, and varying the firm's risk aversion coefficient, λ from 0 to 7 in steps of 1, for each of the four scenarios of interest with respect to the availability of closure (never, 20% of the time, always, every 5 years).

The first notable difference with the risk-neutral case (Goumagias et al., 2012) is that the optimal degree of tax-evasion, $[u]_1$, for $\lambda > 0$ was *not* constant. That is, in every case, our DQN-based approach converged to a state-dependent (static) policy which achieved a higher average utility than would have been possible using any constant value for $[u]_1 \in [0, 1]$ (meaning that the same value of $[u]_1$ would be used regardless of which state we were in). See Tab. 2 for a comparison in the case where $\lambda = 2.6$ (we have chosen this particular value because it will be of special interest in Sec. 5.4.1 - similar results hold for different values of λ).

In terms of the four tax policies under consideration, we observe from Tab. 2 that - as in the risk-neutral case - the firm obtains a higher maximum discounted utility when the closure option is offered more frequently or more predictably. This implies that, from the point of view of government, the tax revenue collected is highest when the closure option is never offered at all. We will have more to say about this in Sec. 5.6.

Regarding the use of closure by the firm ($[u]_2$) we found that, for the tax-parameters currently in use, if the closure option is always offered then the firm must always take advantage of it (so that it is never audited). If the option is offered stochastically or every five years, then it is optimal for the firm to use it *unless* the firm has *just* been audited (this being a departure from the optimal risk-neutral policy). With respect to the level of tax-evasion, $[u]_1$, the fact that the optimal policy is not constant makes it difficult to characterize it in a "compact" way, especially when closure is offered stochastically or periodically. We will discuss ways of exploring the structure of $[u]_1$ later in this Section.

5.4. The effect of risk aversion on tax evasion

To gain insight into the firm's behavior we plotted the *average* $[u]_1$ over the course of the firm's lifetime against the firm's risk-aversion coefficient, λ. Fig. 2 shows the rate at which the average level of tax evasion ($[u]_1$) declines as the firm becomes more risk-averse, for each of the four scenarios regarding the availability of closure, where for each value of λ there were 100 episodes executed with 250 time steps each. The

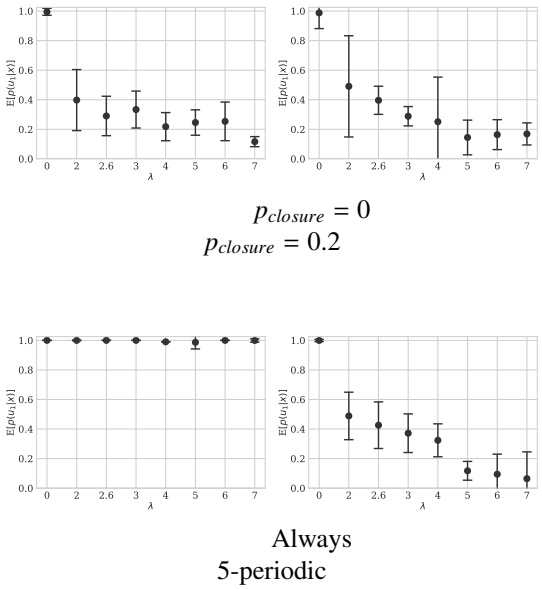


Figure 2: Average level of tax-evasion over the firm's evolution (100 episodes of 250 time steps each). Dots represent mean values, bars indicate ± one standard deviation.

approximate nature of our approach comes through in the fact that in the case where closure is never offered (Fig. 2 - top left), there are times where the average level of tax-evasion increases as λ (the firm's risk aversion) increases, although we expect the opposite to occur. There is, however, a clear downward trend in the vast majority of cases showing that as the firm becomes more

Closure Option	Max. discounted utility (average $[u]_1$)	Max. discounted utility with constant $[u]_1$
Never	$-1.91474\cdot10^{-2}$ (0.29)	$-1.98007\cdot10^{-2}$ (0.21)
$p_{closure} = 0.2$	$-1.87780\cdot10^{-2}$ (0.40)	$-1.94671\cdot10^{-2}$ (0.31)
Always	$-1.40147\cdot10^{-2}$ (1)	$-1.40147\cdot10^{-2}$ (1)
5-periodic	$-1.86345\cdot10^{-2}$ (0.43)	$-1.89893\cdot10^{-2}$ (0.37)

Table 2: Long-term discounted expected utilities for a risk-averse ($\lambda = 2.6$) firm: maximum achieved vs. maximum under the best *constant* $[u]_1$. The numbers in parentheses indicate the time-average value of $[u]_1$ leading to the maximum expected utility, and the optimum *constant* $[u]_1$, respectively.

risk averse (higher λ) the firm becomes more "honest" on average. It is also worth mentioning that it is not trivial to obtain high numerical precision with an approximation method such as ours when the utility function is highly nonlinear (i.e., in our case, very steep near zero where the firm would find itself if it had to pay a penalty at audit time, and relatively "flat" for values of income associated with non-audit states). One possible solution for learning value-functions over different reward "scales" is offered in van Hasselt et al. (2016b); however, the implementation is complex, hence we opted for reward clipping as discussed in Sec. 5.1.

5.4.1. Calculating the risk aversion of Greek firms

In Fig. 2 we included data points for $\lambda = 2.6$ on the horizontal axis. That value of the risk-aversion coefficient is significant because (see Fig. 2 top-right) it leads to a 40% average tax-evasion on behalf of the firm. It was identified by numerical experimentation, essentially using bisection on λ to make the average $[u]_1 = 0.4$. As we have mentioned before, the 40% level is reported in the literature as the estimated tax-evasion level in Greece (Artavanis et al., 2016), and so our approach allows us to estimate the risk aversion coefficient of the average Greek firm (or to re-estimate it for all or a subset of firms, as newer empirical data becomes available).

5.5. Exploring the optimal policy for a representative firm ($\lambda = 2.6$)

As we have seen, the firm's optimal policy is not constant in three of the four closure availability scenarios (the exception is the case where closure is *always* available, where it is always best to conceal all profit). Because of the complexity of the problem and the large number of states (15×101^5), it is difficult to represent or even visualize the optimal policy in a compact form. We have thus attempted to gain insight by examining the statistics of $[u]_1$ and $[u]_2$ and by using decision trees, as well as various projections of the state-to-decision mapping encoded in the DQN that are of practical relevance because they reveal how the tax evasion level is related to i) the tax status of the firm (i.e., how many years since its last audit or closure), and ii) the amounts that the firm has previously concealed but are still within the statute of limitations in the event of an audit.

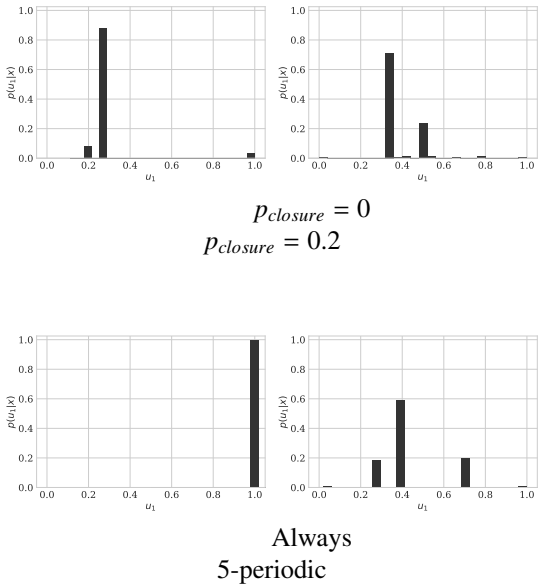


Figure 3: Histograms of the firm's level of tax-evasion throughout its lifetime (100 episodes, 250 steps per episode).

Fig. 3 shows the frequency histograms of the firm's optimal level of tax-evasion over 25000 state-decision pair samples (obtained from our trained DQN, over 100 episodes where the firm was allowed to evolve for 250 steps, as previously mentioned). We observe that there is no variability in the case where closure is always available (the firm always uses the closure option and conceals as much profit as possible). In the cases where the option is offered stochastically or periodically there is more significant variability in the optimal $[u]_1$ (Fig. 3, top row, and bottom-right), although we observe that the set of values for $[u]_1$ used by the DQN is sparse.

To gain insight into just how the values observed

in the histograms depend on the firm's state, we used decision-tree classifiers. Fitting a decision tree to the outputs of the network is a commonly-used approach for discovering patterns in the learned policy. We opted for a shallow decision tree (depth = 3) to the same 25000 outputs $[u]_1$, with a high threshold for splitting (10^{-4}). We kept the tree classifier "naive" in order to be able to gain high-level intuition on the decision policy's structure.

Fig. 4 illustrates the trees obtained for the cases of $p_{closure} = 0$ and 0.2. In the tree nodes, the binary s_i stand for the firm's tax status in terms of the i-th element of S (see state space description following Eq. 1), e.g., $s_5 = 0$ means that the firm's tax status is *not* the fifth element of $\mathcal{S}$, so that the firm is *not* being audited for its last five tax years; h_i denotes the i-th element of the firm's tax history vector h, i.e., the amount of profit it concealed $5 - i + 1$ years ago; $c \in \{0, 1\}$ denotes whether closure is available to the firm or not; and *samples* denotes the number of samples (out of the 25000 total) to which each case applied. The decision tree for $p_{closure}$=0

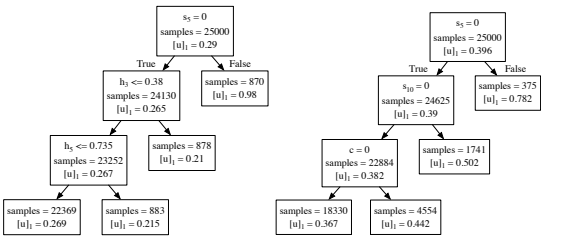


Figure 4: Decision trees analyzing the firm's levels of tax-evasion $[u]_1$ (left: closure never available, right: closure available with a 20% probability) using samples from 100 episodes of 250 steps each. Data were fit to purposely shallow (depth-3) decision trees with a high threshold for splitting, in order to glean information on the high-level structure of the policy.

(Fig. 4-left) indicates that if closure is never offered, the firm opts for very high tax evasion ($[u]_1 = 0.98$, top-right leaf of the tree) only immediately after audits that "cover" the last 5 years. The remainder of the time, the firm almost always conceals 27% of its profit (bottom-left leaf of the decision tree), and any deviations from that value depend mainly on its history h of tax evasion (e.g., whether 3 years ago it concealed more or less than 38% of its profit - see left "child" of the tree's root node).

When $p_{closure} = 0.2$ (Fig. 4-right), the firm again uses a high $[u]_1 = 0.78$ immediately after (rare) audits; for the majority of its time it uses two tax evasion levels, $[u]_1 = 0.44$ or $[u]_1 = 0.37$ depending on whether closure is ($c = 1$) or is not ($c = 0$) available, respectively. For the 5-periodic closure scenario, the classifier (not shown) indicated that when the firm is 3 or 4 years away from the next closure, it uses a near-average $[u]_1 \approx 0.38$. If the closure option is less than 3 years away, and the firm has been recently audited (< 5 years ago), then its tax-evasion goes up to $[u]_1 \approx 0.67$.

To glean additional information on the structure of the DQN policy, we looked for patterns in the tax evasion decisions based on i) the tax status of the firm (i.e., whether it is being audited, using the closure option, or left to evolve with 1-5 years since its last audit or closure, as detailed in Sec. 3.2), and ii) the cumulative tax evasion "stored" in the firm's history (h_k) within the 5-year statute of limitations, this representing a kind of "amount at risk" that the firm would be liable for if it were to be audited.

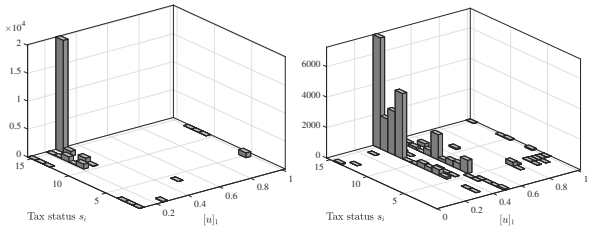


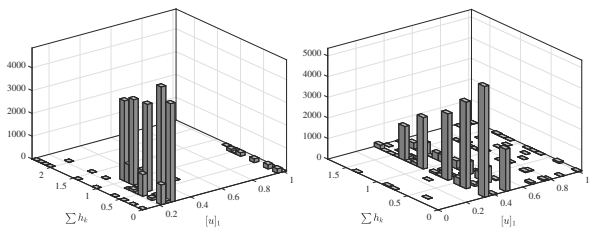
Figure 5: Histograms showing the distribution of 25000 tax evasion decision samples ($[u]_1$) from 100 episodes of 250 steps each. Axes on the horizontal plane correspond to the level of tax evasion (0 to 1) and the integer-coded tax status of the firm (1-15, as explained in Sec. 3.2).

Fig. 5 shows histograms of the firm's decisions according to level of tax evasion ($[u]_1$) and the tax status of the firm (shown as an integer between 1 and 15 representing states in $\mathcal{S}$, as per Sec. 3.2). In the left histogram, where closure is never available, we observe that the firm spends most of its time in the tax status 15 (which corresponds to the firm having been unaudited for 5 or more years) and its level of tax evasion is near 0.28 (this matches the decision tree analysis above). Also noteworthy is the fact that the firm consistently uses $[u]_1 = 1$ when its tax status is 5 (the firm being audited for its last 5 tax filings).

In the right histogram of Fig. 5, the closure option is available with probability 0.2, and if we were to sum over the tax status axis we would obtain the top-right histogram of Fig. 3. The firm generally uses a higher level of tax evasion ($[u]_1 \approx 0.35 - 0.5$). The broader spread of the samples over the tax status axis compared with the previous case (closure never available) indicates that the firm uses the option when it can, thereby

"erasing" any tax evasion history and thus finds itself more often in a tax status of 5-10 (corresponding to closure being used for the last 1-5 tax filings of the firm) or 11-15 (the firm being unaudited for 1-5 years in the past).

Besides grouping the firm's decisions by tax status, we examined how the firm behaves based on the part of its state, h_k, which contains its past tax evasion decisions (up to five) which are still within the statute of limitations (see Sec. 3.2). Because we have quantized $[u]_1$ in steps of 0.01, and because of the structure of h_k as the firm evolves via Eq. 1, it is difficult to visualize the firm's policy over that entire set. It is instructive, however, to consider the sum of the elements of h_k (which is proportional to the total amount the firm has failed to disclose) as a proxy variable for the amount at risk if it the firm were to be audited, and examine how it affects tax evasion by the firm. We expect that a "good" policy would reduce tax evasion ($[u]_1$) when that sum increases, which is precisely what happens. Fig. 6 shows the histograms of the firm's level of tax evasion and sum of its past decisions (up to five or up to the last time it was audited or used the closure option, whichever is smaller). In the left histogram, where closure is never available, we observe that although the firm conceals approximately 30% of its profit most of the time, it sometimes decides to be completely dishonest with $[u]_1$ at 1, when the amount it is potentially "on the hook" for ($\sum h_k$) is small (between 0 and 1.2) but becomes more honest (with $[u]_1$ at 0 or 0.2) when that amount is larger.



$p_{closure} = 0$ $p_{closure} = 0.2$

Figure 6: Histograms showing the distribution of 25000 tax evasion decision samples ($[u]_1$) from 100 episodes of 250 steps each. Axes on the horizontal plane correspond to the level of tax evasion (0 to 1) and the sum of the past history of tax evasion decisions, as explained in Sec. 3.2).

In the right histogram of Fig. 6, the closure option is available with probability 0.2, and there are occurrences of $[u]_1 = 1$ throughout the range of values for $\sum h_k$. This is explained by the fact that the usage of closure allows the firm to "wipe the slate clean" so that it is less deterred by the fact that it has accumulated a history of tax evasion. The downward trend present in the bars near $[u]_1 = 0.35$ is because by using the closure option whenever possible (thereby zeroing out h_k), it is more likely for the firm to find itself with a lower value of $\sum h_k$.

5.6. Tax policy implications

With an eye towards making policy recommendations for the "canonical" firm ($\lambda = 2.6$) we observe that, based on the results of Sec. 5.3, the more frequently the closure option is offered by the government, the higher the firm's expected utility (see left column in Tab. 2), and - correspondingly - the lower the amount of tax-revenue collected. Thus, it appears that government should avoid using this type of tax amnesty because it encourages tax evasion, and instead reinforce auditing mechanisms.

Also, the analysis of the DQN policy in Sec. 5.5 suggests ways in which the tax authority could re-allocate auditing resources towards firms which are in states associated with the highest tax evasion. In particular, under the current regime, most auditing resources are devoted to firms which have not been audited for five years and thus have past tax filings which are about to pass beyond the statute of limitations. The histograms and decision tree analysis of the firm's policy shows that tax evasion is high immediately after an audit, suggesting that the audit probabilities should be distributed more "evenly" on $\mathcal{S}$, to improve the chance of catching tax evaders that were audited just one year ago.

Finally, Fig. 2 gives guidance for the expected reduction in tax evasion as the firm's risk aversion increases. Of course, it is not easy to directly influence firm's attitudes to make them more risk averse. However, the relationship between average $[u]_1$ and λ provides an opportunity for optimizing the allocation of auditing resources among various categories of firms (grouped, for example, by size or sector of economic activity), with fewer audits for the very risk-averse, more for those who are less so, once each group's risk aversion coefficients are estimated (this can be done empirically by examining tax audits to measure tax evasion within each group, and estimating that group's λ as we did in Sec. 5.4.1).

6. Conclusions

This work is part of a research program whose aim is to provide governments with quantitative tools which can be used to combat tax evasion and guide tax policy. A prerequisite for the design of effective policies is to be able to understand, in quantitative terms, the behavior of tax evaders. Towards that end, we addressed the

problem of determining the behavior expected of a self-interested risk-averse firm which aims to maximize its long-term revenues, in a tax system whose features include tax rates, random audits, penalties for tax evasion and occasional tax amnesties. The practical importance of the problem is significant: solving it allows one to estimate tax revenues, to identify measures and parameter values that make self-interested entities behave more honestly, and to gauge the effectiveness of current or planned tax policies.

The dynamics of the firm's (stochastic) evolution, combined with the rules of the tax system and the non-linearity of the firm's reward function (owing to the fact that the firm is generally risk-averse), give rise to a stochastic optimal decision problem in which the associated Bellman equation is difficult to solve using exact methods. To address that challenge, we made use of recent developments in function approximation and neural networks and constructed a Deep Q-learning Network (DQN) which "learns" the optimal firm policy. The neural network was trained to "store" the firm's optimal long-term revenues, given a starting state and decision. DQN was used to efficiently "learn" the optimal firm decisions through simulations of the firm's state evolution.

The DQN approach was first validated by setting our model to the special case of risk neutrality and comparing the results thus obtained (optimal policy and long-term firm revenues) to the exact solution computed via DP (Goumagias et al., 2012). We subsequently demonstrated that we can compute the firm's optimal policy and corresponding tax revenues for the government in the "full" model which includes both risk-aversion (i.e., non-linearity in the reward function) and the tax amnesty ("closure") option. We note that, in our particular case, Deep Learning was successful in approximating the firm's reward function and finding its optimal decisions where other approximation methods failed to converge (we experimented extensively with Approximate Dynamic Programming, various implementations of Q-learning and SARSA algorithms, and neural network architectures which served as function approximators).

One of the contributions made possible by our approach is that it can be used to infer the risk aversion coefficient of a typical taxpayer from empirical data, and thus subsequently evaluate that taxpayer's reactions under various scenarios of tax amnesty availability, or other parameter change (i.e., increase in the audit rates or penalties). Using Greece as a case study, we estimated the risk aversion coefficient of the average firm to be approximately $\lambda = 2.6$, based on empirical evidence that puts the level of the Greek "hidden" economy at approximately 40% (Artavanis et al., 2016). We also compared tax revenues for a series of policies used there; our results provide evidence against the use of tax amnesties as tax revenue collection tools, even within economies with persistent and endemic tax evasion, as we there is a negative relationship between the predictability (or indeed existence) of tax amnesties and tax revenue. Although we have used Greece as a case study here, in part for the sake of concreteness, the proposed approach is adaptable to different taxation schemes and can easily be "tuned" to reflect the values of various tax-parameters, such as audit rates, which are known to the government.

Opportunities for further work include the use of the very recent sample-efficient actor-critic algorithm with experience replay (Wang et al., 2017), which could enable stable learning in continuous action spaces (without having to discretize the firm's decisions); efficient reward scaling, to handle reward values across many orders of magnitude similarly to van Hasselt et al. (2016b); and the use of Recurrent Q-Learning to possibly reduce some state features, e.g., the firm's behavior in the past five-year window.

An interesting (and massive) computational study which has now been made possible in light of the present work, involves recording the effects of altering the various tax parameters on the behavior of the firm, so that one could compute the "degree of honesty" of the firm as a function of the parameters, in the spirit of the maps given in Goumagias et al. (2012).

Finally, we also envision extensions of this work with learning models that generalize over different values of the tax rate r or the risk aversion coefficient λ (instead of having to be trained separately for particular values), or that also optimize selected model parameters simultaneously with the firm's decisions. Although some parameters, such as λ, are generally considered exogenous in forming the firms' risk preferences, optimizing others, especially the tax rate and penalty factor would be of particular interest for the purposes of maximizing tax revenue.

7. References

Allingham, M. G. and Sandmo, A. (1972). Income tax evasion: a theoretical analysis. *J. Public Economics*, 1(3-4):323–338.

Alm, J. and Beck, W. (1990). Tax amnesties and tax revenues. *Pub. Fin. Rev.*, 18(4):433–453.

Alm, J. and Rath, D. M. (1998). Tax policy analysis: the introduction of a russian tax amnesty. Technical report, Georgia State University, Andrew Young School of Policy Studies.

Andreoni, J., Erard, B., and Feinstein, J. (1998). Tax compliance. *J. Econ. Lit.*, 36(2):818–860.

Artavanis, N., Morse, A., and Tsoutsoura, M. (2016). Measuring income tax evasion using bank credit: Evidence from Greece. *The Quarterly Journal of Economics*, 131(2):739–798.

Baldry, J. C. (1979). Tax evasion and labour supply. *Economics Letters*, 3(1):53–56.

Bayer, R. C., Oberhofer, H., and Winner, H. (2015). The occurrence of tax amnesties. *J. Public Economics*, 125:70–82.

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

Bornstein, C. T. and Rosenhead, J. (1990). The role of operational research in less developed countries: A critical approach. *European Journal of Operational Research*, 49(2):156–178.

Clotfelter, C. T. (1983). Tax evasion and tax rates: An analysis of individual returns. *The Review of Economics and Statistics*, pages 363–373.

Cowell, F. A. (1981). Taxation and labour supply with risky activities. *Economica*, 48(192):365–379.

Crane, S. E. and Nourzad, F. (1986). Inflation and tax evasion: An empirical analysis. *The Review of Economics and Statistics*, pages 217–223.

Das-Gupta, A. and Mookherjee, D. (1995). *Tax amnesties in India: an empirical evaluation*. Boston University, Institute for Economic Development.

Fleming, M. H., Roman, J., and Farrell, G. (2000). The shadow economy. *Journal of International Affairs*, 53(2):387–409.

Foerster, J. N., Assael, Y. M., de Freitas, N., and Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In *NIPS*.

Gao, S. and Xu, D. (2009). Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering. *Expert Systems with Applications*, 36(2):1493–1504.

Garrido, N. and Mittone, L. (2012). Tax evasion behavior using finite automata: Experiments in chile and italy. *Expert Systems with Applications*, 39(5):5584–5592.

Gosavi, A. (2004). Reinforcement learning for long-run average cost. *European Journal of Operational Research*, 155(3):654–674.

Goumagias, N., Hristu-Varsakelis, D., and Saraidaris, A. (2012). A decision support model for tax revenue collection in greece. *Decision Support Systems*, 53(1):76–96.

Hellenic Ministry of Finance (2004). Law N.3259/2004 (POL.1034/2005) (in Greek).

Hellenic Ministry of Finance (2008). Law N.3697/2008 (POL.1130/2008) (in Greek).

Hellenic Ministry of Finance (2015). Article 7, Par. 2,4, Law N.4337/2015 (POL.4337/2015) (in Greek).

Hokamp, S. and Pickhardt, M. (2010). Income tax evasion in a society of heterogeneous agents–evidence from an agent-based model. *International Economic Journal*, 24(4):541–553.

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krauss, C., Do, X. A., and Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, 259(2):689–702.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Leung, M. K., Xiong, H. Y., Lee, L. J., and Frey, B. J. (2014). Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12):i121–i129.

Lin, L. (1993). *Reinforcement Learning for Robots Using Neural Networks*. PhD thesis, Carnegie Mellon University, Pittsburgh.

Markellos, R. N., Psychoyios, D., and Schneider, F. (2016). Sovereign debt markets in light of the shadow economy. *European Journal of Operational Research*, 252(1):220–231.

Martinez-Vazquez, J. and Rider, M. (2005). Multiple modes of tax evasion: theory and evidence. *National Tax Journal*, pages 51–76.

Mikolov, T., Deoras, A., Povey, D., Burget, L., and Černockỳ, J. (2011). Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 196–201. IEEE.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Narasimhan, K., Kulkarni, T., and Barzilay, R. (2015). Language understanding for text-based games using deep reinforcement learning. *arXiv preprint arXiv:1506.08941*.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 701–710. ACM.

Pickhardt, M. and Seibold, G. (2014). Income tax evasion dynamics: Evidence from an agent-based econophysics model. *J. Economic Psychology*, 40:147–160.

Ronao, C. A. and Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, 59:235–244.

Ross, J. M. and Buckwalter, N. D. (2013). Strategic tax planning for state tax amnesties evidence from eligibility period restrictions. *Public Finance Review*, 41(3):275–301.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Singh, S. P. (1994). Reinforcement learning algorithms for average-payoff markovian decision processes. In *AAAI*, volume 94, pages 700–705.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Tadepalli, P. and Ok, D. (1996). Scaling up average reward reinforcement learning by approximating the domain models and the value function. In *ICML*, pages 471–479.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3):185–202.

Tsitsiklis, J. N. and Van Roy, B. (1996). Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1-3):59–94.

van Hasselt, H., Guez, A., and Silver, D. (2016a). Deep reinforcement learning with double q-learning. In *AAAI*.

van Hasselt, H. P., Guez, A., Hessel, M., Mnih, V., and Silver, D. (2016b). Learning values across many orders of magnitude. In *NIPS*, pages 4287–4295.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. (2017). Sample efficient actor-critic with experience replay. In *ICLR*.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*.

Wheeler, R. and Narendra, K. (1986). Decentralized learning in finite markov chains. *IEEE Transactions on Automatic Control*, 31(6):519–526.

Yitzhaki, S. (1974). Income tax evasion: A theoretical analysis. *J. Pub. Econ.*, 3(2):201–202.

Appendix A. State dynamics

The parameters of the basic state equation 1 are as in Goumagias et al. (2012) but are also given here for the purposes of review:

$$A = \begin{bmatrix} 0 & & \\ & 0 & \\ & & H \end{bmatrix},\ H = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},\ B = \begin{bmatrix} \begin{pmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \end{bmatrix},\ n_k = \begin{bmatrix} \omega_k \\ \epsilon_k \\ 0_{5\times1} \end{bmatrix} \tag{A.1}$$

The scalar ϵ_k corresponds to the government determining whether to offer the closure option; this occurs with some fixed probability, p_o, each year so that:

$$\Pr(\epsilon_k = 1) = \begin{cases} p_o & if\ i = 1 \quad \text{(option available)} \\ 1 - p_o & if\ i = 2 \quad \text{(option not available)}. \end{cases} \tag{A.2}$$

The scalar $\omega_k \in \{1, ..., |S|\}$ is a random variable corresponding to the transitions that the firm undergoes in S (e.g., tax audits) depend on its current state and its decision $[u_k]_2$ to accept or reject the closure option (if offered):

$$\Pr\left(\omega_\kappa = i | x_k = [j,\ q,\ h_k^T]^T, [u_k]_2 = m\right) = P_{\text{qij}}(m), i, j \in \{1, \dots, 15\},\ q \in \{1, 2\} \tag{A.3}$$

where we use $1, ..., 15$ as labels for states in S. For q and u fixed, $P_{\text{qij}}(m)$ forms a Markov matrix governing the firm's transitions in S:

$$P_{\text{qij}}(m) = \begin{cases} [M_{\text{no}}]_{\text{ij}} & if\ q = 2, \quad \forall\ m \quad \text{(no option)} \\ [M_a]_{\text{ij}} & if\ q = 1, \quad m = 1 \quad \text{(option taken)} \\ [M_d]_{\text{ij}} & if\ q = 1, \quad m = 2 \quad \text{(option declined)} \end{cases} \tag{A.4}$$

and the M_{no}, M_a, and M_d are as in Goumagias et al. (2012) but are also given in Appendix B for the purposes of review.

Appendix B. Markov transition matrices

$$M_{\text{no}} = \begin{bmatrix}
0.0025 & 0.0025 & 0.0025 & 0.0025 & 0.0025 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.0025 & 0.0025 & 0.0025 & 0.0025 & 0.0025 & 0.0025 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0025 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.002 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.9975 & 0.9975 & 0.9975 & 0.9975 & 0.9975 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.9975 & 0.9975 & 0.9975 & 0.9975 & 0.9975 & 0.9975 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9975 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.997 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

Table B.3: Transition Probabilities M_{no} : Closure is not available

$$M_a = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

Table B.4: Transition Probabilities M_a : Closure is available and the firm decides to use it.

$$M_d = \begin{bmatrix}
0.0075 & 0.0075 & 0.0075 & 0.0075 & 0.0075 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.0075 & 0.0075 & 0.0075 & 0.0075 & 0.0075 & 0.0075 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0075 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.007 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.9925 & 0.9925 & 0.9925 & 0.9925 & 0.9925 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.9925 & 0.9925 & 0.9925 & 0.9925 & 0.9925 & 0.9925 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9925 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.992 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

Table B.5: Transition Probabilities M_d : Closure is available and the firm decides not to use it.