*mathematics*

MDPI

*Article*

# A Markovian Mechanism of Proportional Resource Allocation in the Incentive Model as a Dynamic Stochastic Inverse Stackelberg Game

**Grigory Belyavsky, Natalya Danilova and Guennady Ougolnitsky ***

Vorovich Institute of Mathematics, Mechanics and Computer Sciences, Southern Federal University, ul. Milchakova 8A, Rostov-on-Don 344090, Russia; beliavsky@hotmail.com (G.B.); danilova198686@mail.ru (N.D.)
* Correspondence: ougoln@mail.ru

Received: 4 July 2018; Accepted: 27 July 2018; Published: 30 July 2018


**Abstract:** This paper considers resource allocation among producers (agents) in the case where the Principal knows nothing about their cost functions while the agents have Markovian awareness about his/her strategies. We use a dynamic setup of the stochastic inverse Stackelberg game as the model. We suggest an algorithm for solving this game based on *Q*-learning. The associated Bellman equations contain functions of one variable for the Principal and also for the agents. The new results are illustrated by numerical examples.

**Keywords:** dynamic inverse Stackelberg game; incentives; online learning; resource allocation

## 1. Introduction

Stackelberg games date back to the monograph [1]. The original setup includes two players, Leader (Principal) and Follower (Agent). The Leader makes the first move by choosing his/her strategy and informing the Follower of it. After that, the Follower seeks for an optimal response by maximizing his/her payoff function. There are Stackelberg games, in which the Leader has a constant strategy, and inverse Stackelberg games, in which the Leader's strategy is a function of the Follower's actions (feedback control mechanism). Stackelberg games with several Leaders and/or Followers are considered later. Inverse Stackelberg games in the static and dynamic setups are discussed in the surveys [2,3].

Inverse Stackelberg games provide a mathematical formalization for the incentive problem. The Principal designs a feedback control mechanism stimulating the agents to choose actions that are most beneficial for him/her. In this paper, we suggest a method for solving such problems in the dynamic setup under incomplete information about the agents' behavior.

Below the problem is stated so that the associated Bellman equations contain functions of one variable for the Principal and also for the agents. In addition, the agents become independent players as soon as they receive the information from the Principal. The latter chooses his/her behavior based on the response of the players. The agents make their decisions by predicting the Principal's behavior. An elementary implementation of this principle is the Markov principle, i.e., at a time $t$ the Principal uses the agents' response at the time $(t-1)$ to design his/her behavior. In turn, the agents observe the behavioral history of the Principal in order to predict his/her choice at the time $(t+1)$. The equilibrium calculation method for the Stackelberg game relies on online learning (more specifically, the *Q*-learning procedure) and recursive statistical estimation. The *Q*-learning procedure is mostlyused for solving the statistical dynamic programming problem, with the calculation of the *Q*-function. As a rule, the *Q*-function depends on two variables, the phase and control variables; so, the *Q*-function is defined on the Cartesian product of finite sets. The method has slow convergence,

and the rate of convergence essentially depends on the number of elements in the definitional domain of the arguments. Therefore, the Stackelberg game in one of the two setups considered below is stated so that the resulting $Q$-function depends on a single argument. In the other setup, the slow $Q$-learning procedure is replaced by a faster one-dimensional maximization algorithm for a concave function of one variable. In both setups, the agents involve recursive statistics. Thus, in comparison to the standard $Q$-learning procedure, the suggested algorithms are expected to guarantee a faster convergence to the equilibrium of this game.

The main contribution of this paper is as follows:

1. We have developed fast converging algorithms to calculate the solution of the dynamic inverse Stackelberg game without sufficient information about the agents.
2. The suggested algorithms can be considered as numerical methods for solving the corresponding static inverse Stackelberg game without sufficient information about the payoff functions of the agents.
3. This game-theoretic model has been applied for optimal resource allocation among producers in the case of insufficient information about their cost functions.

The remainder of this paper is organized accordingly. In Section 2, we provide a survey of the existing publications on this subject. In Section 3, we discuss the static incentive problem as an inverse Stackelberg game and also describe its dynamic extension in two setups. In Section 4, we present the results of calculations using numerical examples for both setups of the game. Finally, in Section 5, we give some concluding remarks and outline for future research.

## 2. Related Work

Resource allocation mechanisms in the static setup are studied in contract theory [4] and also in control of organizational systems [5]. Here, the main concern of the investigators is to design strategy-proof mechanisms [6,7]. More specifically, by assumption, the Principal does not know the exact characteristics of the agents and the latter can use this fact for strategic manipulation (information distortion for their own benefit), see [8]. A possible way to eliminate manipulation was described in the book [5]; the author suggested a strategy-proof direct resource allocation mechanism. Such an approach relies on the hypothesis that the optimal control is obtained by solving the static inverse Stackelberg game in which the Principal knows the goals of all agents. The paper [9] considered the discrete-time incentive model with Markovian dynamics and discounted payoff function on the infinite planning horizon. As demonstrated here, the approximate Stackelberg solution can be found by solving an optimal control problem with the difference between the controller's income and executor's cost as the optimality criterion.

Static and dynamic inverse Stackelberg games were surveyed in [2,3,10]. Also, note [11–15] among the earlier publications. The paper [16] considered a game-theoretic setup of incentive problems. Linear-quadratic inverse Stackelberg games were studied in [17,18]. Applications of these models were described in [19–22]. The authors [23] provided a survey of hierarchical games with application to marketing.

In particular importance, previous research suggested no common methods for solving inverse Stackelberg games. Meanwhile, the paper [24] proved the theorem on the $\varepsilon$-optimal guaranteeing strategy in the static inverse Stackelberg game, which reduces the maximin problem with bound variables to nonlinear programming problems with independent variables; as a result, calculations were considerably simplified. In [25,26], this approach was extended to dynamic inverse Stackelberg games. The corresponding theorem actually reduces constrained maximin calculation over complex functional spaces to the calculation of multiple maximins over finite-dimensional spaces.

The authors [27] analyzed a dynamic modification of the proportional resource allocation mechanism. Their suggested approach to the game-theoretic modeling of resource allocation in the hierarchical Principal-agent system possesses the following features.

(1) Resource dynamics are explicitly described as the phase variable depending on the Principal's control. The control function can be non-differentiable; in this case, the dynamic equation is interpreted in terms of the Lebesgue-Stieltjes integral.
(2) The Principal's control has smooth variations, which is formalized using the Lipschitz property of the control function. This assumption seems natural for the majority of real organizational and economic systems.
(3) The Principal allocates resources among the agents proportionally to their actions, which stimulates the latter to choose more intensive plans.
(4) This hypothesis is used to develop a genetic algorithm for calculating the Principal's optimal strategy with a non-uniform partition of the time interval.

Evolutionary modeling and genetic algorithms were described in the monographs [28,29]. The paper [30] presented a hybrid learning procedure for artificial neural networks. The authors [31] proposed a genetic algorithm for solving the Germeier game with one Follower and the control function that satisfies the Lipschitz condition. Genetic algorithms for solving Stackelberg games were also considered in [32,33].

This paper is focused on the case in which the Principal has insufficient information about the goals of all agents while the latter does not know the Principal's strategy for the whole duration of the game, merely its history (without loss of generality, the awareness structure will be considered Markovian). The problem statement involves statistical estimation and reinforcement learning, including the *Q*-learning [34,35]. Reinforcement learning was used for calculating Stackelberg equilibria in [36,37]. Particularly, as noted in [36], this algorithm has instability in the case of several Followers, especially if the latter calculates the Nash equilibrium for solving their problems. In this paper, we suggest a Stackelberg game-based model, for which there exists a stable algorithm.

## 3. Model

### *3.1. Static Setup and Dynamic Generalization*

Consider a single Principal and $M$ agents controlled by him/her. The static incentive model as an inverse Stackelberg game has the form

$$\psi(x) - \sum_{i=1}^{M} \phi_i(x,\gamma) \to \max$$
$$\phi_i(x,\gamma) - f_i(x) \to \max,\ i = 1,\dots,M$$

Here $\psi(x)$ denotes the Principal's income, a concave increasing function such that $\psi(0) = 0$; $f_i(x)$ is the cost of agent $i$, a convex increasing function such that $f_i(0) = 0$; $\phi_i(x,\gamma)$ gives the incentive of agent $i$, i.e., the compensation of his/her cost by the Principal; $i = 1, \dots, M$; finally, $\gamma$ indicates the Principal's control strategy. Then the optimal incentive mechanism has the form [5]

$$\phi_i^*(x,\gamma) = \begin{cases} f_i(x_i^*, x_{-i}) \text{ if } x_i = x_i^*, \\ 0 \text{ otherwise } (i = 1,\dots,M), \end{cases}$$

where $x^* \in Arg\max_x \left[\psi(x) - \sum_{i=1}^{M} f_i(x)\right]$.

However, this solution may be impossible to calculate directly due to insufficient information about the cost functions $f_i(x)$ that is available to the Principal. So, we will suggest a computational scheme based on solving a dynamic stochastic inverse Stackelberg game.

Assume the Principal stimulates the activity of all agents by allocating resources among them. Their activity is described by a vector sequence $x^M(t) = (x_i(t))_{i=1}^M$, $t = 0,1,\dots$ At each time $t$, the Principal chooses a control strategy $\gamma(t)$, and each agent $i$ obtains a corresponding part $\phi_i(x_i(t-1),\gamma(t))$, $i = 1,\dots,M$, of the resource. This resource allocation mechanism is selected

because the Principal does not know the cost of all agents; while choosing a local control strategy $\gamma(t)$, he/she may rely just on the available history $x^M(0), \ldots, x^M(t-1)$. This paper considers the case of Markovian processes: at each time $t$, the available information about the preceding value $x(t-1)$ is used only. The Principal cannot inform the agents about his/her control strategies for the whole duration of the game and the agents cannot calculate these strategies. Therefore, the agents (like the Principal) should be guided by the available information at the time $t$, i.e., by the sequence $\gamma(1), \ldots, \gamma(t)$. Below we consider the two setups of the dynamic inverse Stackelberg game as a resource allocation model.

### *3.2. Dynamic Setup 1*

The first model consists in the following. The set of admissible influences on agents (the set of all Principal's scenarios) is finite. The agents suppose that $\gamma(1), \ldots, \gamma(t)$ is a segment of a Markov sequence defined in the stochastic basis $\mathrm{B}^\gamma = \left\langle \Omega^\gamma, [F^\gamma(t)]_{t\geq 0}, F^\gamma(\infty), G^\gamma \right\rangle$, $F^\gamma(t) = \sigma(\gamma(1), \ldots, \gamma(t))$, i.e., the minimal sigma-subalgebra supplemented with the events of zero probability. The agents solve their problems with the discounted payoff functions $E_{G_y^\gamma} \sum\limits_{t=1}^{\infty} \beta^t [\phi_i(x_i(t-1), \gamma(t)) - f_i(x_i(t))]$, where $0 < \beta < 1$ and $E_{G_y^\gamma}(.)$ denotes the conditional expectation operator with the probability measure induced by the sequence $\gamma$ given $\gamma(0) = y$.

Consider the problem of agent $i$:

$$E_{Q_y^\gamma} \sum_{t=1}^{\infty} \beta^t [\phi_i(x_i(t-1), \gamma(t)) - f_i(x_i(t))] \to \max$$

This problem is solved by dynamic programming; the associated Bellman functions satisfy the equations

$$V_i(x, y) = \max_z \left[ \phi_i(x, y) - f_i(z) + \beta \int\limits_{-\infty}^{\infty} V_i(z, u) q^\gamma(du/y) \right]. \tag{1}$$

In addition, $x_i(0) = x$, $\gamma(0) = y$. The measure $q^\gamma(du/y)$ in the integrand expression is the transition core of the Markov sequence $\gamma$. Write the Bellman function in the form $V_i(x, y) = \phi_i(x, y) + W_i(y)$. As a result, we obtain the following Fredholm integral equation of the second kind in the unknown function $W_i(y)$:

$$W_i(y) = \beta \int\limits_{-\infty}^{\infty} W_i(u) q^\gamma(du/y) + \max_z \left[ -f_i(z) + \beta \int\limits_{-\infty}^{\infty} \phi(z, u) q^\gamma(du/y) \right]. \tag{2}$$

So, the existence of a unique solution for the Bellman equation is equivalent to the existence of a unique solution for the Fredholm integral equation of the second kind [38]. Consider the space $B(-\infty, \infty)$ of all bounded functions with the norm $\sup\limits_x |f(x)|$ and also the operator $Gf = \int\limits_{-\infty}^{\infty} f(u) q^\gamma(du/y)$ over this space. Obviously, $\|G\| \leq 1$. If the function $\phi_i(y) = \max\limits_z \left[ -f_i(z) + \beta \int\limits_{-\infty}^{\infty} \phi_i(z, u) q^\gamma(du/y) \right]$ is bounded above, then Equation (2) has a unique solution. For the function $\phi_i(y)$ to be bounded, a sufficient condition is that the functions $\phi_i(z, u) - f_i(z)$ are bounded above for all $u$. Moreover, if the functions $f_i(x)$ are convex and $\phi_i(z, u)$ are convex in the variable $z$ for any $u$, then there exists a unique solution of the problem $z_i^*(y) = \operatorname{argmax}\limits_z \left[ -f_i(z) + \beta \int\limits_{-\infty}^{\infty} \phi(z, u) q^\gamma(du/y) \right]$.

The optimal activity of the agents satisfies the equality

$$x_i^*(t) = z_i^*(\gamma(t)). \tag{3}$$

What is of fundamental importance, the optimal behavior of all agents in this model depends on the Principal's local control strategy $\gamma(t)$ only—-in no way on the preceding values of the sequence. Therefore, the Principal's problem is to calculate

$$\max_{\gamma(t)}\left[\psi\left(\sum_{i=1}^{M} z_i^*(\gamma(t))\right) - \sum_{i=1}^{M}\phi(x_i(t-1),\gamma(t))\right]. \tag{4}$$

The Principal does not know the relationship (3), while the agents do not know the transition core of the Markov sequence $\gamma$. In our case, this core is defined by the transition probability matrix because a Markov chain has a finite set of admissible states. As mentioned earlier, the Principal and agents can observe the response of each other. So, the Principal learns by observing the response of agents to his/her actions while the agents learn by observing the Principal's response to their actions. The problem becomes much easier in comparison with the general $Q$-learning procedure for solving the dynamic inverse Stackelberg games that wereconsidered in [36]: for the agents, it is required to approximate the transition probability matrix; for the Principal, to maximize the function of one variable. Statistical estimation of a transition probability matrix and maximization of a function of one variable are not so computationally intensive as calculation of the $Q$-functions for the Principal and agents. Let $\Gamma = \{\gamma^1,\ldots,\gamma^r\}$ and denote by $P$ the transition probability matrix. The maximum likelihood estimate $P_t$ of this matrix is designed as follows. Consider a matrix sequence $G$ such that $G_t(i,j) = G_{t-1}(i,j) + I_{\{\gamma_{t-1}=\gamma^i,\gamma_t=\gamma^j\}}$, where $I_{\{S\}}$ means the indicator of a set $S$. The initial value is $G_0 = 0$. The maximum likelihood estimate $P_t$ has the form $P_t(i,j) = \frac{G_t(i,j)}{\sum\limits_k G_t(i,k)} I_{\{\sum\limits_k G_t(i,k)\neq 0\}}$. Let $\gamma(t) = \gamma^j$. Then $\max\limits_z\left[-f_i(z) + \beta\int\limits_{-\infty}^{\infty}\phi_i(z,u)q^\gamma(du/y)\right]$ can be written as the approximate problem

$$\max_z\left[-f_i(z) + \beta\sum_k P_t(j,k)\phi_i\left(z,\gamma^k\right)\right]. \tag{5}$$

If the function $F_t^i(z) = -f_i(z) + \beta\sum\limits_k P_t(j,k)\phi_i\left(z,\gamma^k\right)$ is concave and bounded above, then problem (5) has the unique solution $x_i^*(t) = \operatorname{argmax} F_t^i(z)$. Then the Principal observes $R\left(\gamma^j\right)$, i.e., the response of agents to the control strategy $\gamma^j$: $R\left(\gamma^j\right) = \psi\left(\sum\limits_{i=1}^{M} x_i^*(t)\right) - \sum\limits_{i=1}^{M}\phi_i\left(x_i(t-1),\gamma^j\right)$. After that, the Principal modifies the $Q$-function using the reinforcement learning algorithm

$$Q_{t+1}\left(\gamma^j\right) = Q_t\left(\gamma^j\right) + h_t\left(R\left(\gamma^j\right) - Q_t\left(\gamma^j\right)\right), \tag{6}$$

and chooses the next control strategy for the agents based on the probability distribution $P_{t+1}^l = \left\{p_{t+1}^l(\gamma^1),\ldots,p_{t+1}^l(\gamma^r)\right\}$. The probabilities can be calculated by the Boltzmann scheme, which is used in annealing [36]. This is a random search algorithm for maximization (minimization) of generally non-differentiable functions $f(x)$. The idea consists in comparing the current solution $x_t$ with a randomly chosen one $y_t$ in a small neighborhood of the former. Transition to the randomly chosen solution occurs if $f(y_t) > f(x_t)$ ($f(y_t) < f(x_t)$). If this condition fails, then transition to a next

random value $y_t$ is performed with the probability $p_t = \exp\left(-\frac{f(x_t)-f(y_t)}{T_t}\right)$ $\left(p_t = \exp\left(\frac{f(x_t)-f(y_t)}{T_t}\right)\right)$. Therefore, it is natural to choose the next value $\gamma$ with the probability

$$p_{t+1}^l\left(\gamma^j\right) = \exp\left(Q_{t+1}^l\left(\gamma^j\right)/T_{t+1}\right)/\sum_k \exp\left(Q_{t+1}^l\left(\gamma^k\right)/T_{t+1}\right). \quad (7)$$

In formula (7) and also above, the parameter $T$—"temperature"—adjusts the degree of randomness for control strategies. This parameter is decreasing from a given maximal value to a given minimal value, e.g., $T_{t+1} = \delta T_t$, $0 < \delta < 1$. The initial condition is $Q_0^l \equiv 0$. The values $h_t$ regulating the amplitude of variations are supposed to satisfy the standard conditions $\sum\limits_{t=1}^{\infty} h_t = \infty$, $\sum\limits_{t=1}^{\infty} h_t^2 < \infty$. They guarantee the almost sure convergence to the optimal function $Q$ that is close to $R$ if each element of the set $\Gamma$ appears an infinite number of times in the course of learning.

Thus, the suggested algorithm consists of the following iterative operations:

1. calculation of the maximum likelihood estimate for the transition probability matrix;
2. calculation of the next value of the $Q$-function.

Once again, we underline an important advantage of the suggested algorithm: the $Q$-function depends on a single argument only. The first operation is to calculate the maximum likelihood estimate instead of a next value of the $Q$-function, which guarantees faster convergence of the algorithm against its counterparts in which the $Q$-function depends on two arguments and a next value of $Q$ is calculated at the first and second stages. Also, note that the maximum likelihood estimates are consistent and asymptotically efficient, i.e., they all use available information and obeys the normal distribution in asymptotics.

### *3.3. Dynamic Setup 2*

In the second model of this game, the agents assume that the Principal makes "no sudden moves" in control choice (see Postulate 2 from [27]). So, they describe the Principal's behavior by

$$\gamma(t) = \gamma(t-1)(1+\varepsilon_t). \quad (8)$$

The sequence $\varepsilon_t$ consists of independent identically distributed random elements with the mass concentrated near the origin. For the second model, $\max\limits_z \left[-f_i(z) + \beta \int\limits_{-\infty}^{\infty} \phi_i(z,u) q^{\gamma}(du/y)\right]$ can be written as $\max\limits_z[-f_i(z) + \beta E_\varepsilon \phi(z, y(1+\varepsilon))]$. Consider the second term and apply the Taylor expansion up to the second order inclusive, following the standard approach of stochastic calculus (e.g., the Ito formula). As a result, we obtain the optimization problem

$$\max_z \left[-f_i(z) + \beta\left(\phi_i(z,y) + \frac{\partial \phi_i(z,y)}{\partial y} y E_\varepsilon \varepsilon + \frac{1}{2}\frac{\partial^2 \phi_i(z,y)}{\partial y^2} y^2 E_\varepsilon \varepsilon^2\right)\right], \quad (9)$$

provided that the second derivative exists. If for all $y$ the goal functions of the agents are concave and bounded, then there exist unique solutions $z_i^*(y)$ for the agents' problems (9). The sample means $a_t = \frac{1}{t}\sum\limits_{i=1}^{t} \frac{\Delta\gamma(i)}{\gamma_{i-1}}$, $b_t = \frac{1}{t}\sum\limits_{i=1}^{t}\left(\frac{\Delta\gamma(i)}{\gamma_{i-1}}\right)^2$ for the moments of distribution of $\gamma$ are consistent estimates for the moments in the right-hand side of Formula (9). Therefore, the agents solve the problems $\max\limits_z\left[-f_i(z) + \beta\left(\phi_i(z,\gamma(t)) + \frac{\partial \phi_i(z,\gamma(t))}{\partial y}\gamma(t)a_t + \frac{1}{2}\frac{\partial^2 \phi_i(z,\gamma(t))}{\partial y^2}\gamma(t)^2 b_t\right)\right]$, while the Principal can use any maximization method of a concave function of one variable without derivatives. If the function $R(x)$ is concave and bounded above, then the consistent sample moments and the convergent one-dimensional search procedure guarantee that this algorithm converges to the equilibrium of the Stackelberg game in probability.

Thus, the learning procedure for the second model consists of the following iterative operations:

1. calculation of the estimates for the first and second moments of distribution;
2. calculation of the next approximation $\gamma$.

In comparison with the classical $Q$-learning [34,35], this approach does not calculate the $Q$-function and is based on estimating the first and second moments of distribution and one-dimensional maximization of a function of one variable, which forms an obvious advantage. In comparison with the previous algorithm (see Section 3.2), it does not need the preliminary analysis of the problem to find the Principal's behavioral scenarios (the set $\Gamma$).

## 4. Examples and Numerical Calculations

Consider an example in which $\psi(x) = \sqrt{x}$, $\phi_i(x,y) = xy$, and $f_i(x) = \mu_i x^2$. For the first model, problem (5) takes the form: $\max\limits_z \left[-\mu_i z^2 + \beta z \sum\limits_k P_t(j,k)\gamma^k\right]$, with the solution $x_i^*(t) = \frac{1}{2\mu_i}\beta \sum\limits_k P_t(j,k)\gamma^k$. The response is $R(\gamma^j) = \sqrt{\sum\limits_{i=1}^{M} x_i^*(t)} - \gamma^j \sum\limits_{i=1}^{M} x_i^*(t-1)$.

For the second model, problem (9) takes the form $\max\limits_z \left[-\mu_i z^2 + \beta z(\gamma(t) + a_t)\right]$, with the solution $x_i^*(\gamma(t)) = \frac{\beta(\gamma(t)+a_t)}{2\mu_i}$, where $a_t = \frac{1}{t}\sum\limits_{i=1}^{t} \frac{\Delta\gamma(i)}{\gamma(i-1)}$. The response $R(\gamma(t))$ is the same as above.

For this problem, the Stackelberg equilibrium satisfies the equalities $x_i^* = \frac{\beta\gamma}{2\mu_i}$, $\gamma^* = \left(\frac{1}{2\sqrt{2\beta\sum\limits_{i=1}^{M}\frac{1}{\mu_i}}}\right)^{2/3}$.

Choose the following numerical parameters for trial calculations:
Then the equilibrium values are

$$\gamma^* = 0.4524029361,\ x_1^* = 0.2035813212,\ x_2^* = 0.1017906606.$$

Consider two examples as follows.

In the first example, the set $\Gamma = \{0.4, 0.45, 0.5\}$ contains a value close to the equilibrium in position 2. Our calculations were performed in Maple. The algorithm yielded the following results. The calculated values of the $Q$-function are $Q = \text{vector}(0.02789054257,\ 0.4073228841,\ 0.09271767315)$. The maximal value of the $Q$-function is 0.4073228841 (position 2), which corresponds to the value $\gamma = 0.45$ from the set $\Gamma$. The calculated transition probability matrix has the form $P = \text{matrix}([0,\ 3/4,\ 1/4],\ [1/91,\ 87/91,\ 3/91],\ [3/4,\ 1/4,\ 0])$. The maximal element of this matrix, which is close to 1, stands at the junction of column 2 and row 2. In other words, the most probable prefix of the chain is 0.45, 0.45, ..., which also corresponds to the value $\gamma = 0.45$ from the set $\Gamma$. The calculated values $x_1 = 0.2029945055$, $x_2 = 0.1014972528$ are close to the equilibrium.

In the second example, the set $\Gamma = \{0.2, 0.4, 0.6\}$ has no values close to the equilibrium. The calculated values of the $Q$-function are $Q = \text{vector}(0.1254077349,\ 0.4028884358,\ 0.05489325521)$. The maximal value of the $Q$-function is 0.4028884358 (position 2), which corresponds to the value $\gamma = 0.4$ from the set $\Gamma$. The calculated transition probability matrix has the form $P = \text{matrix}([1/5,\ 3/5,\ 1/5],\ [1/90,\ 43/45,\ 1/30],\ [3/4,\ 1/4,\ 0])$. The maximal element of this matrix, which is close to 1, also stands at the junction of column 2 and row 2. The calculated values $x_1 = 0.1820000000$, $x_2 = 0.09100000000$ slightly vary from the equilibrium.

The second model. The results yielded by the algorithm for the second model are presented in Tables 1 and 2.

**Table 1.** The results yielded by the algorithm.

| | |
|---|---|
| $M$ | 2 |
| $\beta$ | 0.9 |
| $\mu_1$ | 1 |
| $\mu_2$ | 2 |

**Table 2.** The results of numerical calculations.

| $\gamma$ | $x_1$ | $x_2$ | $R$ |
|---|---|---|---|
| 0.35 | 0.1837500000 | 0.09187500000 | 0.2625000000 |
| 0.36 | 0.1778142858 | 0.08890714290 | 0.4172257999 |
| 0.37 | 0.1788773809 | 0.08943869045 | 0.4193054183 |
| 0.38 | 0.1816893340 | 0.09084466700 | 0.4200877844 |
| 0.39 | 0.1852001900 | 0.09260009500 | 0.4207793681 |
| 0.40 | 0.1890599915 | 0.09452999575 | 0.4214115632 |
| 0.41 | 0.1931187782 | 0.09655938910 | 0.4219456888 |
| 0.42 | 0.1973015878 | 0.09865079390 | 0.4223502262 |
| 0.43 | 0.2015667888 | 0.1007833944 | 0.4226042619 |
| 0.44 | 0.2058894150 | 0.1029447075 | 0.4226943805 |
| 0.45 | 0.2102535881 | 0.1051267940 | 0.4226120220 |
| 0.445 | 0.2070930676 | 0.1035465338 | 0.4170062493 |
| 0.4475 | 0.2078141892 | 0.1039070946 | 0.4193088334 |
| 0.45 | 0.2085934568 | 0.1042967284 | 0.4190913001 |
| 0.44875 | 0.2075715326 | 0.1037857663 | 0.4175845248 |
| 0.449375 | 0.2075256146 | 0.1037628073 | 0.4180160874 |

This table has the same notations as before. The obtained results indicate that fast convergence and additional comments are unnecessary.

The second algorithm seems to be preferable if the admissible set $\Gamma$ has no values close to the equilibrium. In our example, the second algorithm demonstrated a faster convergence to the equilibrium than the first. However, for the set $\Gamma$ containing a value close to the equilibrium, the first algorithm yielded more accurate results. At the same time, the second algorithm had a higher accuracy rate for the set $\Gamma$ without such values.

## 5. Conclusions and Future Work

Proportional allocation is the most natural mechanism to distribute resources, which has been approved by the practical control of organizational systems. For this mechanism, the problem of strategy-proofness (protection against manipulation) comes at the forefront because the agents are interested in overrating their real resource demands unknown to the Principal. The static proportional resource allocation mechanism was studied in control of organizational systems (see [5]); a modification of this mechanism that guarantees strategy-proofness was also designed there.

In this paper, we have suggested a dynamic proportional resource allocation mechanism based on learning. We have constructed two stochastic models (setups) of the dynamic inverse Stackelberg game, each guaranteeing the existence of stationary equilibrium. The first model involved the ideology of a finite set of the Principal's behavioral scenarios while the second relied on the natural limits of all Principal's actions. Both models have been illustrated using numerical examples. Each model is associated with an algorithm to find equilibrium in the inverse Stackelberg game.

The experimental results allow us to formulate the following hypothesis. The developed algorithms for solving the dynamic stochastic inverse Stackelberg game can be also used for solving the corresponding static inverse Stackelberg game with insufficient information about the cost functions of all agents. This hypothesis still needs deeper analysis, which will be the subject of futureresearch.

**Author Contributions:** All of the authors have contributed to this work.

**Funding:** This work was supported by the Russian Science Foundation, project No. 17-19-01038.

**Acknowledgments:** We are grateful to the reviewers for careful reading of the manuscript and helpful remarks.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Von Stackelberg, H. *Marktform und Gleichgewicht*; Springer-Verlag: Vienna, Austria, 1934.
2. Olsder, G.J. Phenomena in inverse Stackelberg games, part 1: Static problems. *J. Optim. Theory Appl.* **2009**, *143*, 589–600. [CrossRef]
3. Olsder, G.J. Phenomena in inverse Stackelberg games, part 2: Dynamic problems. *J. Optim. Theory Appl.* **2009**, *143*, 601–618. [CrossRef]
4. Laffont, J.-J.; Martimort, D. *The Theory of Incentives: The Principal-Agent Model*; Princeton University Press: Princeton, NJ, USA, 2002.
5. Novikov, D. *Theory of Control in Organizations*; Nova Science Publishers: New York, NY, USA, 2013.
6. Korgin, N.A. Equivalence and strategy-proofness of non-anonymous priority allotment mechanisms. *Autom. Remote Control* **2016**, *77*, 2065–2079. [CrossRef]
7. Korgin, N.A.; Korepanov, V.O. An efficient solution of the resource allotment problem with the Groves–Ledyard mechanism under transferable utility. *Autom. Remote Control* **2016**, *77*, 914–942. [CrossRef]
8. Nisan, N.; Roughgarden, T.; Tardos, E.; Vazirani, V. (Eds.) *Algorithmic Game Theory*; Cambridge University Press: Cambridge, UK, 2007.
9. Rokhlin, D.B.; Ougolnitsky, G.A. Stackelberg Equilibrium in a Dynamic Stimulation Model with Complete Information. *Autom. Remote Control* **2018**, *79*, 691–702. [CrossRef]
10. Ehtamo, H.; Kitti, M.; Hamalainen, R.P. Recent Studies on Incentive Design Problems in Game Theory and Management Science. In *Advances in Computational Management Science*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2002; Volume 5, Chapter 8; pp. 121–134.
11. Cruz, J.B., Jr. Leader-follower strategies for multilevel systems. *IEEE Trans. Autom. Control* **1978**, *23*, 244–255. [CrossRef]
12. Ho, Y.C. On incentive problems. *Syst. Control Lett.* **1983**, *3*, 63–68. [CrossRef]
13. Ho, Y.-C.; Luh, P.B.; Muralidharan, R. Information structure, Stackelberg games, and incentive controllability. *IEEE Trans. Autom. Control* **1981**, *26*, 454–460.
14. Leitmann, G. On generalized Stackelberg strategies. *J. Optim. Theory Appl.* **1978**, *26*, 637–643. [CrossRef]
15. Simaan, M.; Cruz, J.B., Jr. Additional aspects of the Stackelberg strategy in nonzero-sum games. *J. Optim. Theory Appl.* **1973**, *11*, 613–626. [CrossRef]
16. Ho, Y.-C.; Luh, P.B.; Olsder, G.J. A control-theoretic view on incentives. *Automatica* **1982**, *18*, 167–179. [CrossRef]
17. Ehtamo, H.; Hamalainen, R.P. Construction of optimal affine incentive strategies for linear-quadratic Stackelberg games. In Proceedings of the 24th IEEE Conference on Decision and Control, Fort Lauderdale, FL, USA, 11–13 December 1985; pp. 1093–1098.
18. Tolwinski, B. Closed-loop Stackelberg solution to a multistage linear quadratic game. *J. Optim. Theory Appl.* **1981**, *34*, 485–501. [CrossRef]
19. Behrens, D.A.; Caulkins, J.P.; Feichtinger, G.; Tragler, G. Incentive Stackelberg Strategies for a Dynamic Game on Terrorism. In *Advances in Dynamic Game Theory*; Jørgensen, S., Quincampoix, M., Vincent, T., Eds.; Annals of the International Society of Dynamic Games; Birkhauser: New York, NY, USA; Boston, MA, USA, 2007; Volume 9, pp. 459–486.
20. Luh, P.B.; Ho, Y.-C.; Muralidharan, R. Load adaptive pricing: An emerging tool for electric utilities. *IEEE Trans. Autom. Control* **1982**, *27*, 320–329. [CrossRef]
21. Shen, H.; Basar, T. Incentive-based pricing for network games with complete and incomplete information. In *Advances in Dynamic Game Theory*; Jørgensen, S., Quincampoix, M., Vincent, T., Eds.; Annals of the International Society of Dynamic Games; Birkhauser: New York, NY, USA; Boston, MA, USA, 2007; Volume 9, pp. 431–458.
22. Stankova, K.; Olsder, G.J.; Bliemer, M.C.J. Comparison of different toll policies in the dynamic second-best optimal toll design problem: Case study on a three-link network. *Eur. J. Transp. Infrastruct. Res.* **2009**, *9*, 331–346.

23. He, X.; Prasad, A.; Sethi, S.P.; Gutierrez, G.J. A survey of Stackelberg differential game models in supply chain and marketing channels. *J. Syst. Sci. Syst. Eng.* **2007**, *16*, 385–413. [CrossRef]
24. Germeier, Y.B. On two-person games with a fixed sequence of moves. *Doklady Math.* **1971**, *1989*, 1001–1004. (In Russian)
25. Danil'chenko, T.N.; Kononenko, A.F. Dynamic models of multilevel control systems with information transfer, I. *Prob. Control Inf. Theory* **1984**, *13*, 53–68.
26. Danil'chenko, T.N.; Kononenko, A.F. Dynamic models of multilevel control systems with information transfer, II. *Prob. Control Inf. Theory* **1984**, *13*, 121–139.
27. Belyavsky, G.I.; Danilova, N.V.; Ougolnitsky, G.A. Evolutionary methods for solving dynamic resource allocation problems. *Math. Game Theory Appl.* **2018**. in press. (In Russian)
28. Fogel, D. *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*; IEEE Press: Piscataway, NJ, USA, 2006.
29. Goldberg, D. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*; Kluwer Academic Publishers: Norwell, MA, USA, 2002.
30. Belyavsky, G.I.; Lila, V.B.; Puchkov, E.V. Algorithm and software implementation for hybrid learning method of artificial neural networks. *Softw. Prod. Syst.* **2012**, *4*, 96–101. (In Russian)
31. Belyavsky, G.I.; Danilova, N.V.; Ougolnitsky, G.A. Evolutionary modeling in sustainable management of active systems. *Math. Game Theory Appl.* **2016**, *8*, 14–29.
32. Liu, B. Stackelberg–Nash equilibrium for multilevel programming with multiple followers using genetic algorithms. *Comput. Math. Appl.* **1998**, *36*, 79–89. [CrossRef]
33. Goykhman, M. On self-play computation of equilibrium in poker. *arXiv*, 2018.
34. Sutton, R.S.; Barto, A. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.
35. Watkins, C.J.C.H.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
36. Tharakunnel, K.; Bhattacharyya, S. Single-leader–multiple-follower games with boundedly rational agents. *J. Econ. Dyn. Control* **2009**, *33*, 1593–1603. [CrossRef]
37. Erev, E.; Roth, A. Predicting how people play game: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* **1998**, *88*, 848–881.
38. Eidelman, Y.; Milman, V.; Tsolomitis, A. *Functional Analysis: An Introduction*; AMS: Provence, RI, USA, 2004.