sensors

MDPI

Article

Optimal Greedy Control in Reinforcement Learning

Alexander Gorobtsov [1,2,*], Oleg Sychev [3,*], Yulia Orlova [3], Evgeniy Smirnov [1], Olga Grigoreva [1], Alexander Bochkin [1] and Marina Andreeva [1]

[1] Higher Mathematics Department, Volgograd State Technical University, Lenin Ave, 28, Volgograd 400005, Russia
[2] Mechanical Engineering Research Institute, Russian Academy of Sciences, Maly Kharitonyevsky Pereulok, 4, Moscow 101990, Russia
[3] Software Engineering Department, Volgograd State Technical University, Lenin Ave, 28, Volgograd 400005, Russia
* Correspondence: vm@vstu.ru (A.G.); o_sychev@vstu.ru (O.S.); Tel.: +7-8442-24-84-87

Abstract: We consider the problem of dimensionality reduction of state space in the variational approach to the optimal control problem, in particular, in the reinforcement learning method. The control problem is described by differential algebraic equations consisting of nonlinear differential equations and algebraic constraint equations interconnected with Lagrange multipliers. The proposed method is based on changing the Lagrange multipliers of one subset based on the Lagrange multipliers of another subset. We present examples of the application of the proposed method in robotics and vibration isolation in transport vehicles. The method is implemented in FRUND—a multibody system dynamics software package.

Keywords: optimal control; variational methods; machine learning; reinforcement learning; robotics

Citation: Gorobtsov, A.; Sychev, O.; Orlova, Y.; Smirnov, E.; Grigoreva O.; Bochkin A.; Andreeva M. Optimal Greedy Control in Reinforcement Learning. *Sensors* **2022**, *22*, 8920. https://doi.org/10.3390/s22228920

Academic Editors: Anastasios Doulamis, Nikolaos Doulamis and Athanasios Voulodimos

Received: 30 October 2022
Accepted: 15 November 2022
Published: 18 November 2022

1. Introduction and Related Works

The problem of optimal control is an important scientific problem in various fields of technology, e.g., robotics, vibration damping systems, etc. The exact theoretical solution to this problem can be achieved by using Bellman's dynamic programming method [1] and Pontryagin's maximum principle [2]. However, these methods are limited to low-dimensional equations because of their high computational complexity. Today, various variational formulations of optimal control problems, in particular, reinforcement learning, have become widely used. When using this approach, the control problem is simplified by parametrizing the control function and reducing it to the parametric optimization problem. However, a number of topical control problems (for example, in robotics) still have too high dimensionality to be solved efficiently [3–11]. Some studies (e.g., We et al. [12] and Tu Vu et al. [13]) have investigated the stability problem of perturbed control motion for known referenced motion, which is a much easier task. Our study is aimed at finding the reference motion.

In the general form, the optimal control problem has the following formulation [1]. For the system described by differential equations

$$\mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t) = 0, \tag{1}$$

where $\mathbf{x}(t)$ is the coordinate vector of the entire system with dimension n. We need to find control functions, $\mathbf{u}(t)$, that let us achieve the extreme value of the criterion

$$I = \int_0^T R(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t)dt. \tag{2}$$

As a rule, sign-constant functions are used as the R function. It was previously noted that the exact solution for Equations (1) and (2) using Bellman's dynamic programming

method and Pontryagin's maximum principle was only obtained for several cases of low-dimension tasks [1,2]. The optimal control problem (1 and 2) is transformed into the parametric optimization problem in the reinforcement learning method by using a discrete form of recording the optimization criterion and parameterizing the control function. The optimization criterion is [14]:

$$I = \sum_{i=0}^{i=N} R_i(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t)\gamma^i, \tag{3}$$

where R_i is the value of the criterion function corresponding to the i-th moment of time, and γ is the discount coefficient, which takes a value from 0 to 1. It is assumed that the time interval of control T is divided into N sections. The control function is parameterized on basic functions and takes the form $\mathbf{u}(\mathbf{s}, t)$, where $\mathbf{s}$ is the parameter of the control function. Neural networks, Fourier series, etc., can be used as basic functions [14]. The discount coefficient γ allows the optimality criterion to "weaken" (2). Formula (3) corresponds to a discrete formulation of Bellman's optimal control problem when the discount coefficient is equal to one. The control function is called "greedy" in the reinforcement learning method when it was obtained with a discount coefficient equal to zero. The parameterization of the control function for multidimensional problems leads to high-dimensional optimization problems and also makes the solution dependent on the basic functions on which the control function was interpolated. Moreover, the dependence of the control function on time ties it to time-dependent external disturbances. All this makes developing new methods of solving variational formulations of machine learning problems important.

2. Theoretical Description

Consider the optimal control problem for systems with constraints in the form of algebraic equations. The equations of the state of these systems can be written as

$$\begin{aligned} \mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t) &= 0 \\ \mathbf{Q}(\mathbf{x}, t) &= 0, \end{aligned} \tag{4}$$

where $\mathbf{Q}(\mathbf{x}, t)$ is the constraint equation vector with dimensionality $k \leq n$. For the numerical solution, system (4) is usually used in the form [15–17]

$$\begin{aligned} \mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t) + \mathbf{D}^{\mathbf{T}}\mathbf{p} &= 0 \\ \mathbf{D}\dot{\mathbf{x}} &= \mathbf{h}(\mathbf{x}, t), \end{aligned} \tag{5}$$

where $\mathbf{D}$ is the matrix of coefficients of the constraint equations with dimension $k \times n$, $\mathbf{p}$ is the k-dimensional vector of Lagrange multipliers, and $\mathbf{h}(\mathbf{x}, t)$ is the vector of the right parts of derivatives of the constraint equations. The second equation of system (5) is obtained by differentiating the constraint equations with respect to time. The physical meaning of the Lagrange multipliers for the problems of the dynamics of mechanical systems is the constraint reactions. As the applications considered in this article are related to mechanical systems, the term "constraint reactions" will be used as equal to the term "Lagrange multipliers".

The differential-algebraic system in Equation (5) is widely used in multibody systems (MBS) dynamics software packages for modeling the dynamics of connected systems of bodies [16]. The features of numerical integration (5) related to ensuring stability are considered in [17]. In numerical integration (5), derivatives of coordinates and Lagrange multipliers from the system of linear algebraic equations are found at each integration step according to

$$\begin{pmatrix} \mathbf{M} & \mathbf{D}^{\mathbf{T}} \\ \mathbf{D} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \dot{\mathbf{x}} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{f}^*(\mathbf{x}, t) \\ \mathbf{h}(\mathbf{x}, t) \end{pmatrix}. \tag{6}$$

The solution to a system of linear algebraic equations can be written in the following form

$$\begin{pmatrix} \dot{\mathbf{x}} \\ \mathbf{p} \end{pmatrix} = \mathbf{A}^{-1}\mathbf{b}. \tag{7}$$

Consider the control problem as described in [18]. There is a subset of reactions $\mathbf{p}_1$ in the vector of constraint reactions $\mathbf{p}$ whose elements are numbered from the set K_1; the number of elements in the set K_1 is k_1.Their values are described by functions $\varphi_i(t)$, $i = 1, 2, \ldots, k_1$, or in the matrix form

$$\mathbf{p}_1 = \boldsymbol{\varphi}(t). \tag{8}$$

There is also a subset of k_2 reactions, $\mathbf{p}_2$, from the vector of constraint reactions, $\mathbf{p}$, whose values are taken from subset K_2, which can vary due to changes in the values of unknown functions $h_{2j}(t)$, $j = 1, 2, \ldots, k_2$, which will be called corrective terms. Each reaction from $\mathbf{p}_2$ corresponds to its own constraint equation; the relevant corrective term $h_{2j}(t)$ is added to the right part of this equation. The corrective terms $h_{2j}(t)$ form a column matrix $\mathbf{h}_2$ in the k_2 dimension. Reactions $\mathbf{p}_2$, generally, are control functions, so we will consider k_2 as the number of control functions. The values of reactions $\mathbf{p}_1$, taking into account (7) and the corrective terms, are

$$\mathbf{p}_1 = \mathbf{A}_1^{-1}(\mathbf{b} + \mathbf{h}_2^*(t)), \tag{9}$$

where $\mathbf{A}_1^{-1}$ corresponding to set K_1 is the submatrix of $\mathbf{A}^{-1}$, consisting of the rows $\mathbf{A}^{-1}$ whose numbers belong to K_1. Only the components with numbers from K_2 are non-zero in the column matrix $\mathbf{h}_2^*$ of dimension $n + k$. We assume that $\mathbf{p}_{10} = \mathbf{A}_1^{-1}\mathbf{b}$, $\mathbf{A}_1^{-1}\mathbf{h}_2^* = \mathbf{C}\mathbf{h}_2$, matrix $\mathbf{C}$ contains only columns from matrix $\mathbf{A}_1^{-1}$ with numbers from K_2 and has dimensionality $k_1 \times k_2$. Then (9) takes the form

$$\mathbf{p}_1 = \mathbf{p}_{10} + \mathbf{C}\mathbf{h}_2(t). \tag{10}$$

Taking into account that $\mathbf{p}_1 = \boldsymbol{\varphi}(t)$, from (10), we can obtain the system of linear algebraic equations for determining $\mathbf{h}_2(t)$.

$$\mathbf{C}\mathbf{h}_2(t) = \boldsymbol{\varphi}(t) - \mathbf{p}_{10}. \tag{11}$$

If the system of linear Equation (11) is joint, then it is possible to determine the control functions of $\mathbf{p}_2$ as

$$\mathbf{p}_2 = \mathbf{A}_2^{-1}(\mathbf{b} + \mathbf{h}_2^*(t)), \tag{12}$$

where $\mathbf{A}_2^{-1}$ is the submatrix $\mathbf{A}^{-1}$ corresponding to set K_2, consisting of rows $\mathbf{A}^{-1}$ whose numbers belong to K_2. Given that $\mathbf{p}_{20} = \mathbf{A}_2^{-1}\mathbf{b}$, $\mathbf{A}_2^{-1}\mathbf{h}_2^* = \mathbf{B}\mathbf{h}_2$, matrix $\mathbf{B}$ contains only columns from matrix $\mathbf{A}_2^{-1}$ with numbers from K_2 and has dimension $k_2 \times k_2$. Equation (12) can be rewritten as

$$\mathbf{p}_2 = \mathbf{p}_{20} + \mathbf{B}\mathbf{h}_2(t). \tag{13}$$

Equation (11) gives the values of changes on the right sides of the constraint equations, ensuring the achievement of the desired values of reactions $\mathbf{p}_1$. Since $\mathbf{h}_2(t)$ affects all the variables in the system (4), when integrating the equations of the mathematical model, the accelerations and constraint reactions are calculated from the system with the modified right side as follows

$$\begin{pmatrix} \dot{\mathbf{x}} \\ \mathbf{p} \end{pmatrix} = \mathbf{A}^{-1}(\mathbf{b}+\mathbf{h}_2^*(t)). \tag{14}$$

Equation (11) has a unique solution if $k_1 = k_2$ and matrix $\mathbf{C}$ is non-singular. The properties of matrix $\mathbf{C}$ are determined by the properties of matrix $\mathbf{A}$. The main reason for the singularity of matrix $\mathbf{A}$ is redundant constraints, i.e., linearly dependent rows in matrix $\mathbf{D}$. Redundant constraints can be inherent to the system's structure or introduced on

purpose, for example, in the parallel-structure mechanisms. In the following, it is assumed that the square matrix **A** is non-singular unless otherwise stated.

Case $k_1 = k_2$ is the simplest. Cases $k_1 \neq k_2$ are of greater interest, so we will consider them further.

The systems where $k_1 > k_2$ are commonly called underactuated systems. The analysis of the controlled motion of these systems can be found in [19]. It is difficult to obtain a meaningful solution for such systems within the framework of the considered approach.

Systems where $k_1 > k_2$ are called overactuated. These systems are widespread, for example, in robotics. Their analysis is relevant to our study. We consider the methods of solving the linear system of equations (11) in this case. Matrix **C** of the system is rectangular—with dimensions $k_1 \times k_2$. As already mentioned, matrix **C** is a full-rank matrix.

System (12) can be converted to a system with a square matrix by adding equations. The simplest way of achieving this is by adding linear equations for the corrective terms $\mathbf{h}_2$, i.e., converting (11) to the form

$$\begin{pmatrix} \mathbf{C} \\ \mathbf{V}_1 \end{pmatrix} \mathbf{h}_2(t) = \begin{pmatrix} \boldsymbol{\varphi}(t) - \mathbf{p}_{10} \\ \mathbf{b}_1 \end{pmatrix}, \tag{15}$$

where $\mathbf{V}_1$ is the non-singular matrix of constant terms with dimensionality $(k_2 - k_1) \times k_2$, and $\mathbf{b}_1$ is the column matrix of constant terms on the right side. Only $(k_2 - k_1) \times (k_1 + 1)$ terms are linearly independent in the second equation of the system (15), so matrix $\mathbf{V}_1$ can be represented as

$$\mathbf{V}_1 = \begin{pmatrix} \mathbf{E} & \mathbf{V}_1^* \end{pmatrix}, \tag{16}$$

where **E** is the identity matrix of dimensionality $(k_2 - k_1) \times (k_2 - k_1)$, and $\mathbf{V}_1^*$ is the matrix of arbitrary coefficients of dimensionality $(k_2 - k_1) \times k_1$. This method of reduction to a single solution will be called the method of additional equations for corrective terms.

Additional equations to (11) can be formed by imposing linear connections on the controls. The second equation of system (15) will take the form

$$\mathbf{V}_1 \mathbf{p}_2 = \mathbf{b}_1.$$

Substituting $\mathbf{p}_2 = \mathbf{p}_{20} + \mathbf{B}\mathbf{h}_2(t)$, we will get

$$\mathbf{V}_1 \mathbf{B} \mathbf{h}_2(t) = \mathbf{b}_1 - \mathbf{V}_1 \mathbf{p}_{20}.$$

System (15) is now

$$\begin{pmatrix} \mathbf{C} \\ \mathbf{V}_1 \end{pmatrix} \mathbf{h}_2(t) = \begin{pmatrix} \boldsymbol{\varphi}(t) - \mathbf{p}_{10} \\ \mathbf{b}_1 - \mathbf{V}_1 p_{20} \end{pmatrix}. \tag{17}$$

We call (17) the method of reduction to a single solution by additional equations for controls.

Another way to eliminate the uncertainty of solution (11) is the conditional extremum method. Consider the conditions for the extremum of the expression

$$I = \mathbf{p}_2^T \mathbf{V}_2 \mathbf{p}_2, \tag{18}$$

where $\mathbf{V}_2$ is a diagonal matrix of weights. Therefore, (18) is the weighted sum of the squares of controls. Consider the problem of finding the conditional extremum of expression (18), taking into account the conditions (11). In this case, (18) will be represented as

$$I^* = \mathbf{p}_2^T \mathbf{V}_2 \mathbf{p}_2 + (\mathbf{C}\mathbf{h}_2(t) - \boldsymbol{\varphi}(t) - \mathbf{p}_{10})\boldsymbol{\lambda}, \tag{19}$$

where $\boldsymbol{\lambda}$ is the column matrix of Lagrange multipliers of dimension k_1. Extremum conditions for (20) are

$$\frac{\partial I^*}{\partial h_{2i}} = 0, i = 1, 2, \ldots, k_2, \tag{20}$$

from which we get

$$\begin{pmatrix} \mathbf{B}_1 & \mathbf{C}^T \\ \mathbf{C} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{h}_2 \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_{21} \\ \mathbf{b}_{22} \end{pmatrix}, \tag{21}$$

where $\mathbf{B}_1$ is the matrix of dimension $k_2 \times k_2$ with elements

$$b_{1lm} = 2v_{2mm} \sum_{i=1}^{k_2} b_{im} b_{lm}, l, m = 1, 2, \ldots, k_2, \tag{22}$$

the column matrix $\mathbf{b}_{21}$ of dimension k_2 with elements is

$$b_{21l} = 2 \sum_{i=1}^{k_2} v_{2ii} p_{20i} b_{li}, l = 1, 2, \ldots, k_2, \tag{23}$$

the column matrix $\mathbf{b}_{22}$ of dimension k_1 is

$$\mathbf{b}_{22} = \boldsymbol{\varphi}(t) + \mathbf{p}_{10}. \tag{24}$$

The system of linear equations (21) has the square matrix of coefficients of dimension $k_1 + k_2$ and allows a single solution to be obtained. The method based on the use of (20) will be called the conditional extremum method with constraints in the form of equations of program reactions or simply the conditional extremum method.

This method allows taking into account $k_2 - k_1$ more of the constraint equations. The linear combinations of forces in the actuators (13) can be used as these constraints. In this case, expression (20) will be

$$I^* = \mathbf{p}_2^T \mathbf{V}_2 \mathbf{p}_2 + (\mathbf{C}\mathbf{h}_2(t) - \boldsymbol{\varphi}(t) - \mathbf{p}_{10})\boldsymbol{\lambda} + \mathbf{V}_3(\mathbf{p}_{20} + \mathbf{B}\mathbf{h}_2(t))\boldsymbol{\lambda}_1, \tag{25}$$

where $\mathbf{V}_3$ is the matrix of weights with dimension $k_3 \times k_2, k_3 \leq k_2 - k_1$ and $\boldsymbol{\lambda}_1$ is the corresponding vector of Lagrange multipliers of dimension k_3. The linear system, Equation (21), for the functional (25) will have the following form

$$\begin{pmatrix} \mathbf{B}_1 & \mathbf{C}^T & \mathbf{B}_2^T \\ \mathbf{C} & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_2 & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{h}_2 \\ \boldsymbol{\lambda} \\ \boldsymbol{\lambda}_1 \end{pmatrix} = \begin{pmatrix} \mathbf{b}_{21} \\ \mathbf{b}_{22} \\ \mathbf{b}_{23} \end{pmatrix}, \tag{26}$$

with the matrix $\mathbf{B}_2$ as

$$\mathbf{B}_2 = \mathbf{V}_3 \mathbf{B}, \tag{27}$$

and the matrix $\mathbf{b}_{23}$ as

$$\mathbf{b}_{23} = -\mathbf{V}_3 \mathbf{p}_{20}. \tag{28}$$

The method based on the use of (25) will be called the conditional extremum method with constraints in the form of forces in the actuators. If $k_3 + k_1 = k_2$, system (26) splits into two independent systems. The column matrix $\mathbf{h}_2$ is unambiguously determined from the second and third equations of (26), which are similar to system (17). Function (25) is close to the function of the Karush–Kuhn–Tucker (KKT) method, which is well-known in the theory of nonlinear programming [8]. However, in our case, it is used without any additional conditions that are used in the KKT.

The conditional extremum method makes it possible to reduce the optimal control problem to the parametric optimization problem over the state space of relatively small dimensionality, compared with the direct parameterization of the control functions.

3. Case Studies

The method considered in Section 2 is implemented in the MBS dynamics software FRUND [20]. The examples below are solved using it.

3.1. Inverted Double Pendulum

Consider the flat inverted double pendulum shown in Figure 1. These pendulums are often considered in problems of control synthesis of walking robots [21–32]. The limit on the magnitude of the torque in the pendulum support is a condition for the stability of the robot (i.e., avoiding overturning). The problem is to find the law of change of torque at points B and C with an arbitrary law of motion of point A. Simple usage of the constraint equations in one or two directions at point A leads either to the fixation of the pendulum in its original position or to the fall of the pendulum if one connection is specified in the horizontal direction.

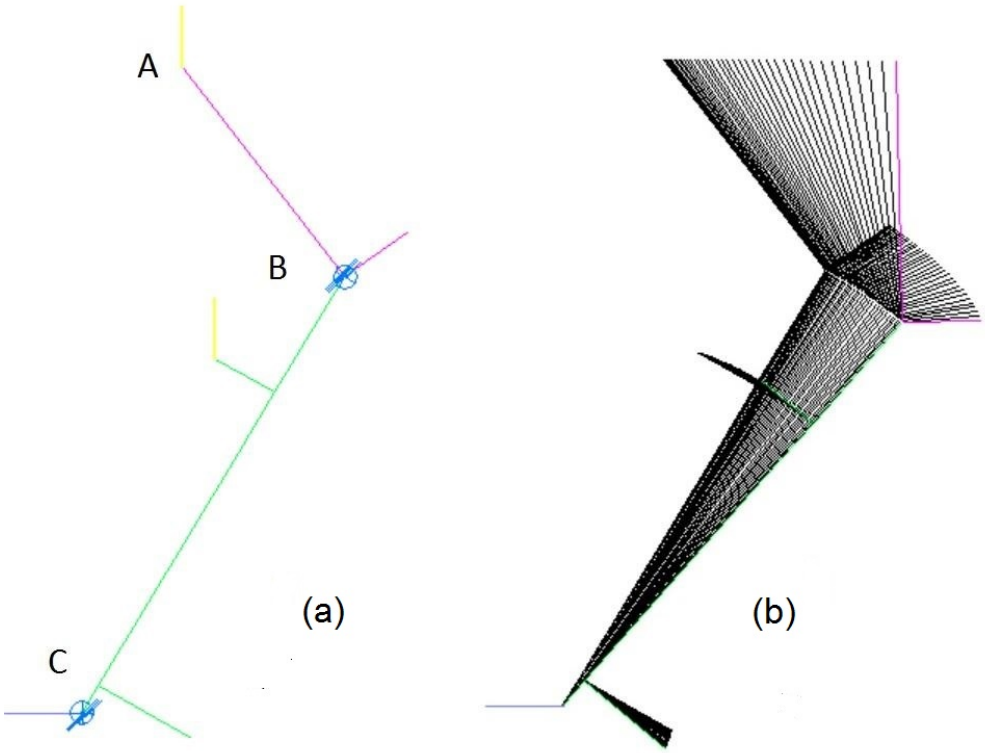

Figure 1. Calculation scheme (**a**) and motion picture (**b**) of the fall of the inverted double pendulum with the condition of horizontal movement of a given point A. A is the point for which reference motion is defined. B and C are pendulum links.

Various options for finding control torques can be considered within the framework of the proposed method. The simplest case is specifying the constraints at point A along the vertical coordinate $Z_A = 0$. The control torque will be found only at point B. The parameters of the dimension that was introduced in Section 2 are $n = 6$, $k = 6$, $k_1 = 1$, $k_2 = 1$; the control torque M_B is found from Equation (13). The motion picture of the pendulum is presented in Figure 1b. The plot of the torque change M_B is presented in Figure 2. During the calculations, it was assumed that the control torque smoothly reaches the program's preset value in 0.05 seconds. The sharp increase in the control torque at the end of the movement is explained by the approach to the singularity position—aligning the pendulum links along the same line (see Figure 1b). Therefore, in this simplest case, the problem of determining the control torque is solved unambiguously.

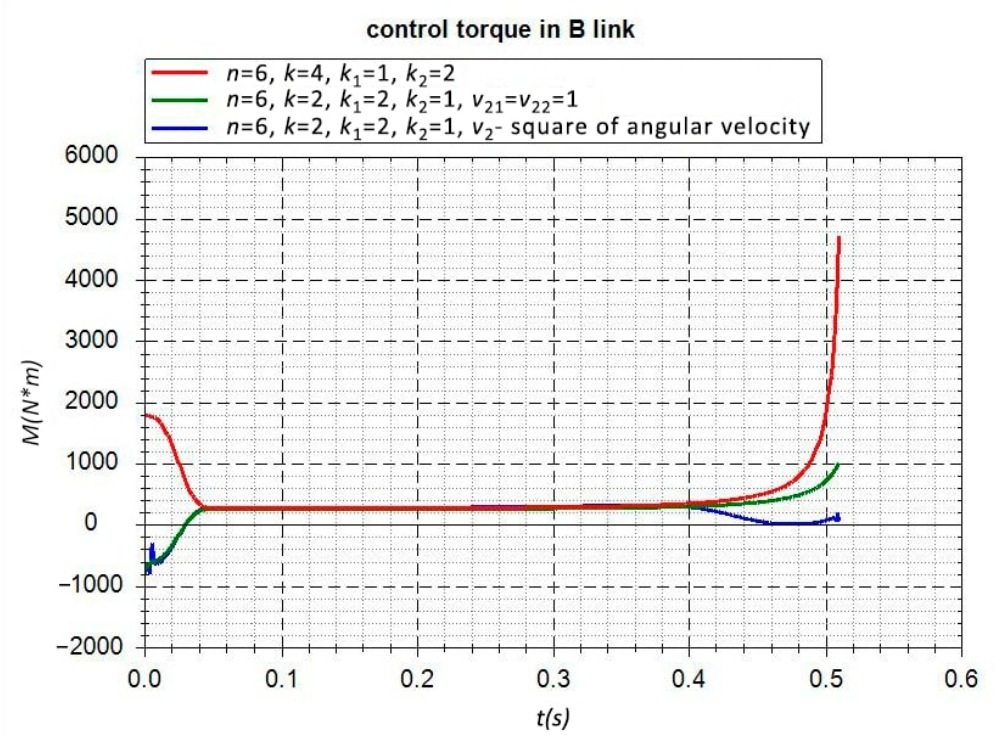

Figure 2. The control torque in link *B* when the horizontal movement of point *A* is free with various control options.

3.2. Spatial Model of Android

Let us consider the spatial motion of a mechanical system on the example of an android robot. The calculation scheme of such a robot is presented in Figure 3. The system parameters are $n = 150$, $k = 144$. The number of actuators in the android structure is 21. The calculation scheme contains two masses with large values of inertial parameters to determine 12 reactions in the contacts of the android's feet with the supporting surface. This method allows for solving some special cases of systems with redundant constraints. In this case, the redundant reactions are six reactions in the feet. Calculations were made for the variant of 8 control drives and restrictions on 6 reactions—$k_1 = 6$, $k_2 = 8$. Control drives are rotation drives working around the transverse axis of the robot in the hinges of its shoulders, hips, knees, and feet—two for each type of hinge. We modeled the displacement of the center of mass of the robot back by 2 cm in 2 s. Vertical reactions and reaction torques were considered unchanged relative to the transverse axis. Horizontal reactions in the feet were calculated from the horizontal inertia forces caused by the movement of the center of mass. As the torque of the reaction was set to be unchanged while the static torque of this reaction increased due because of the movement of the center of mass, this change was compensated for by the movement of the robot's sections, in particular, by the rotation of its arms (see Figure 4). This movement corresponds to the natural reaction of a human when trying to maintain balance without being able to move their legs. The results shown in Figure 4 are obtained by resolving the ambiguity using the conditional extremum method—Equation (21). The squares of angular velocities in the corresponding hinges were taken as weights. During calculations with the same unit weights, the torque compensation occurred solely because of the movement of the body.

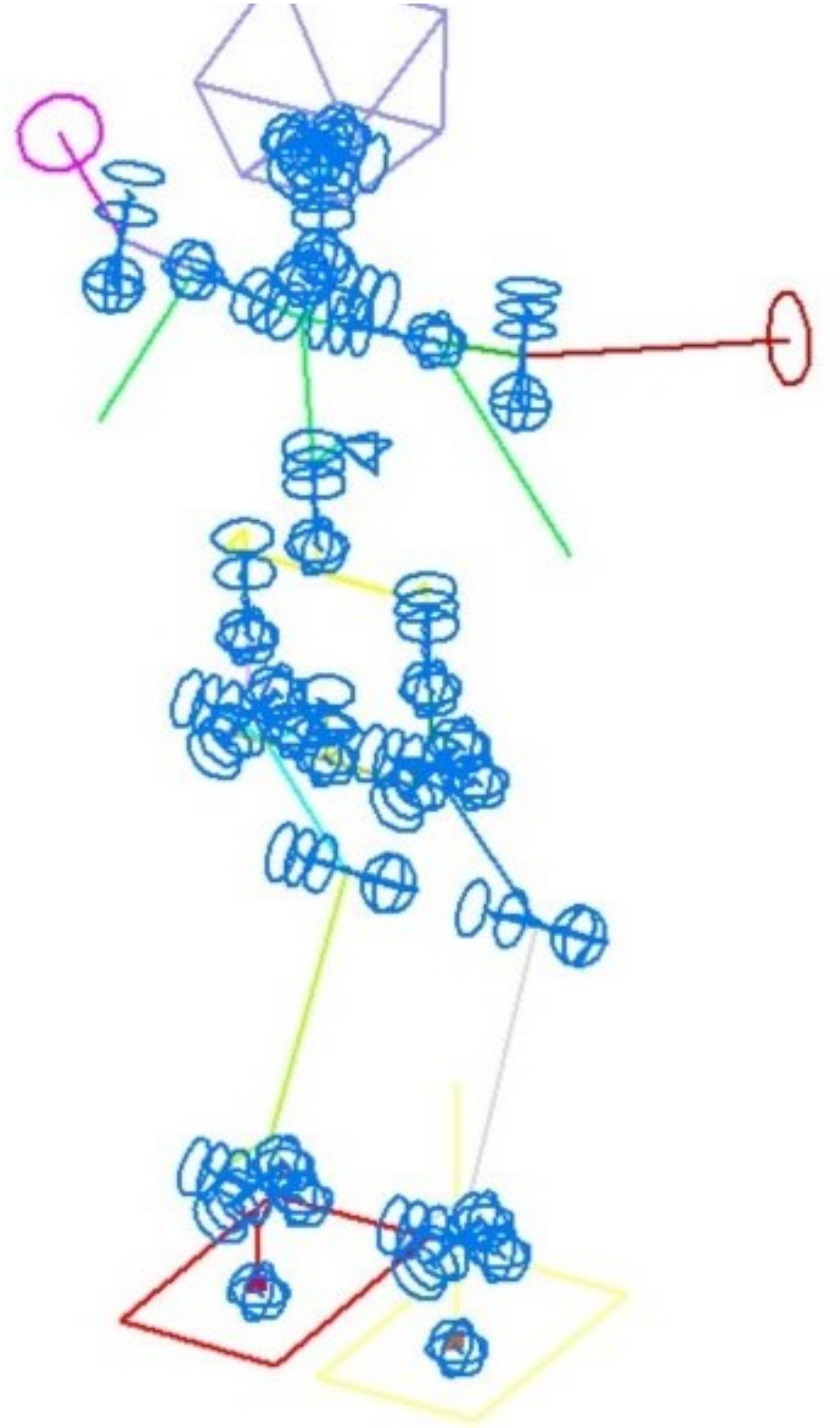

Figure 3. The calculation scheme of an android robot. Blue markers are the links; other colors mark the different bodies of the system.

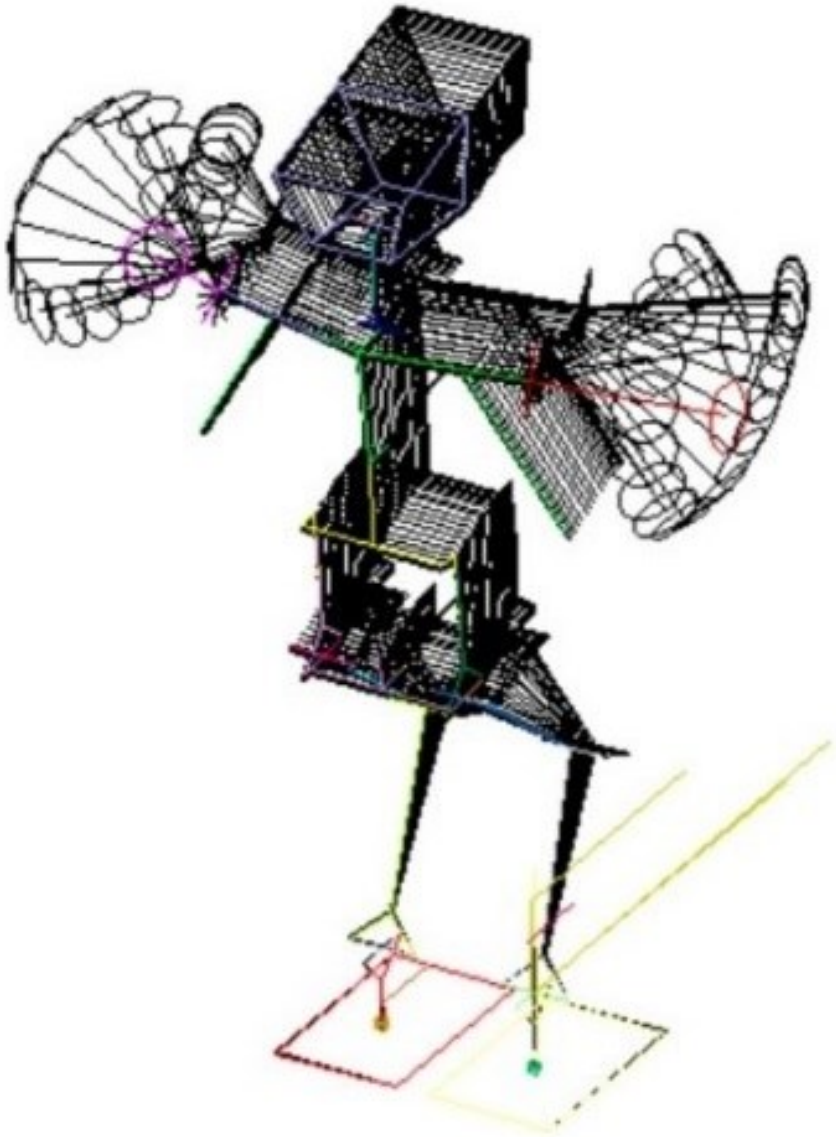

Figure 4. The movement picture of the android's movement when the center of mass is shifted backward while maintaining the magnitude of the reaction torque in the support relative to the transverse axis.

The considered example of controlling an android robot as a multibody mechanical system allows the conclusion that the method is sufficiently versatile and applicable to a wide class of mechanisms to be made, for example, for parallel mechanisms [33–37].

3.3. Optimal Control Problems in the Example of Car Vibrations

Optimization criteria such as (2) are widely used in practical applications of mathematics and mechanics [38–40]. Consider the classical problem of controlling a vibration-isolating system using the car suspension example. Minimization criterion (2) can be presented in the proposed approach as follows

$$J = \int_0^T R^*(\mathbf{p}_1)dt. \tag{29}$$

As R^* is a function of some program values of Lagrange multipliers $\mathbf{p}_1$, which are assigned to desirable functions $\boldsymbol{\varphi}(t)$, the sum of the components of vector $\boldsymbol{\varphi}(t)$ can be used as R^* function. The minimum of functional (29) is achieved, for example, by functions $\boldsymbol{\varphi}(t)$

being equal to zero. The greedy control criterion (3), in this case, will take the following form

$$I = \sum_{i=1}^{i=k_1} \varphi_i(t). \tag{30}$$

The controls for criterion (30) can be found by using (10)–(25). Let us emphasize that the control obtained from criterion (30) is greedy control. However, it is the optimal control as well because it provides the minimum of criterion (29).

Consider the problem of controlling a car's suspension to reduce its vibrations from the impact of the road's micro profile. The existing problem statements can be found in [41–44]. Figure 5 shows the calculation scheme of the mathematical model of the car, which makes simulating its movement along the road irregularities possible. We simulated the movement of the car through a triangular irregularity. Vertical accelerations in the front of the car are presented in Figure 6. In the controlled version, two connections are set—zero vertical movements at two symmetrical points, A and B, in the front of the car body. The $M1$ and $M2$ torques in the two front suspension arms are used as actuators. The dimension parameters are $n = 96, k = 85, k_1 = 2, k_2 = 2$. Using (12), we determined the control torques in the suspension levers, ensuring the movement of points on the frame with zero reactions. Frame accelerations at the considered points are close to zero (see Figure 6). This problem can be considered an example of solving an optimal control problem with the optimization criterion in the form of zero displacements of the selected points of a mechanical system.

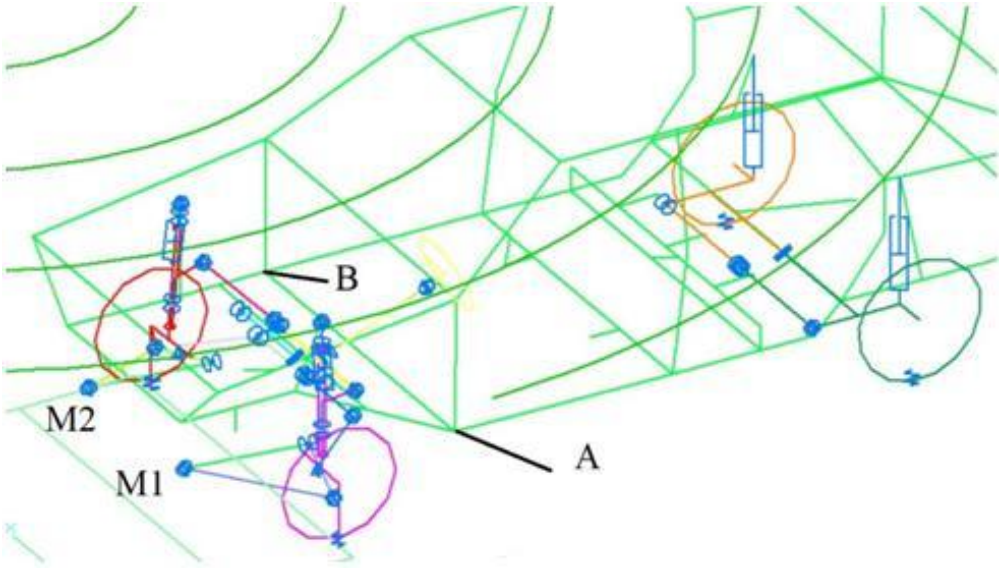

Figure 5. The calculation scheme of a car. A and B is the points whose reference motion is defined. M1 and M2 are the links where the control torques are applied.

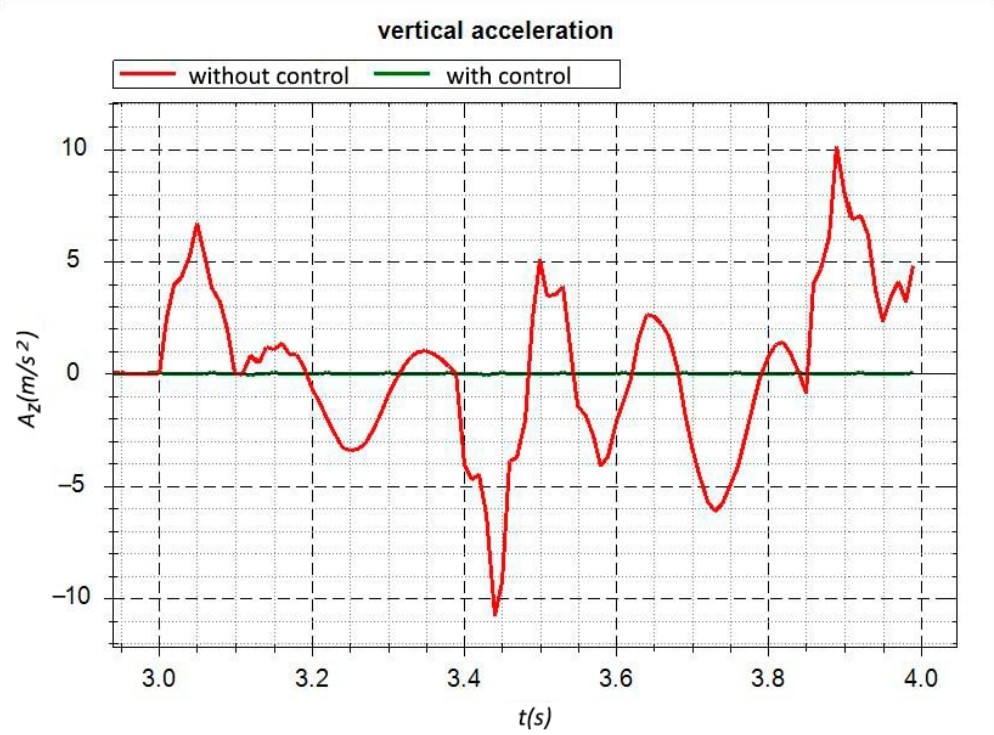

Figure 6. Vertical acceleration in the car's front part.

4. Conclusions

The proposed method of calculating control can be considered a universal theoretical method for solving a wide range of problems related to controlled system dynamics, including the problems of controlling robot manipulators, anthropomorphic and zoomorphic robots, vibration damping problems, etc. The important feature of this method is that it is based on numerical models of machine dynamics, which are widely used in existing computer simulation programs for the dynamics of mechanical systems. The method has no fundamental limitations on the dimensionality of modeled systems and types of nonlinearities.

The evaluation of the proposed method on the described use cases and other test examples proved that computational efficiency has increased for all problems described by DAE (differential-algebraic equations). It was achieved for DAE with a wide range of state dimensions—from 12 to 180 ($k_1 + k_2$) and control dimensions from 1 to 8. The dimensionality of the parameter space is independent of the state dimension and defined only by the number of controls.

The proposed method is a universal theoretical method for the optimal control problem of the systems meeting the following requirements:

- the system is described by DAE (2), which has numerical solutions; constraint Equation (1) is a function of coordinates (holonomic constraints in mechanics);
- the integral object function contains only Lagrange multipliers (29);
- matrix **A** is not singular;
- the linear system in Equation (11) is joint, i.e., it has at least one solution.

The important feature of this method is that it can be considered a kind of machine learning, in particular, reinforcement learning, as a variational formulation of the control problem. The formulation of the proposed method in the form of functionals (20) and (25) corresponds to the so-called "greedy" control [14] in reinforcement learning methods and, at the same time, is the optimal control with the appropriate formulation of integral optimality criteria. From this point of view, the considered method can significantly reduce

the dimensionality of the parameter space, and consequently, increase the computational efficiency of machine learning methods.

The purpose of the presented method is to provide the reference optimal trajectories and controls in the case of the agent having complete knowledge of the environment. The stability problem, controller optimization, and uncertainty model fall beyond the borders of this study. For the tasks of robot control, the standard methods of achieving robustness can be used [45].

This work presents the fundamental theoretical provisions of the method and does not address such issues as control stability and control in systems with singular matrices. These issues are the subject of further research.

Author Contributions: Methodology, A.G.; software, E.S. and M.A.; validation, A.B.; formal analysis, Y.O. and O.S.; investigation, O.G.; writing—original draft preparation, A.G.; writing—review and editing, O.S.; visualization, A.G.; supervision, O.S.; project administration, Y.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code of the developed program FRUND can be accessed freely at svn://dump.vstu.ru/frund.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DAE	differential-algebraic equations
MBS	multibody system
KKT	Karush–Kuhn–Tucker method

References

1. Bellman, R. *Dynamic Programming*; Princeton Landmarks in Mathematics and Physics; Princeton University Press: Princeton, NJ, USA, 2010.
2. Pontryagin, L. *Mathematical Theory of Optimal Processes*; Classics of Soviet Mathematics; Taylor & Francis: Abingdon, England, 1987.
3. Heess, N.; Dhruva, T.; Sriram, S.; Lemmon, J.; Merel, J.; Wayne, G.; Tassa, Y.; Erez, T.; Wang, Z.; Eslami, S.M.A.; et al. Emergence of Locomotion Behaviours in Rich Environments. *arXiv* **2017**, arXiv:1707.02286.
4. Tassa, Y.; Erez, T.; Todorov, E. Synthesis and stabilization of complex behaviors through online trajectory optimization. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, 7–12 October 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 4906–4913. [CrossRef]
5. Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.I.; Abbeel, P. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016.
6. Schulman, J.; Levine, S.; Moritz, P.; Jordan, M.I.; Abbeel, P. Trust Region Policy Optimization. *arXiv* **2015**, arXiv:1502.05477.
7. Duan, Y.; Chen, X.; Houthooft, R.; Schulman, J.; Abbeel, P. Benchmarking Deep Reinforcement Learning for Continuous Control. In *JMLR Workshop and Conference Proceedings, Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York, NY, USA, 19–24 June 2016; JMLR.org; Balcan, M., Weinberger, K.Q., Eds.; Microtome Publishing: Brookline, MA, USA, 2016*; Volume 48; pp. 1329–1338.
8. Tulshyan, R.; Arora, R.; Deb, K.; Dutta, J. Investigating EA solutions for approximate KKT conditions in smooth problems. In Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2010, Portland, OR, USA, 7–11 July 2010; Pelikan, M., Branke, J., Eds.; ACM: New York, NY, USA, 2010; pp. 689–696. [CrossRef]
9. Fu, J.; Li, C.; Teng, X.; Luo, F.; Li, B. Compound Heuristic Information Guided Policy Improvement for Robot Motor Skill Acquisition. *Appl. Sci.* **2020**, *10*, 5346. [CrossRef]
10. Cho, N.J.; Lee, S.H.; Kim, J.B.; Suh, I.H. Learning, Improving, and Generalizing Motor Skills for the Peg-in-Hole Tasks Based on Imitation Learning and Self-Learning. *Appl. Sci.* **2020**, *10*, 2719. [CrossRef]
11. Peters, J.; Schaal, S. Reinforcement learning of motor skills with policy gradients. *Neural Netw.* **2008**, *21*, 682–697. [CrossRef] [PubMed]

12. Wen, G.; Ge, S.S.; Chen, C.L.P.; Tu, F.; Wang, S. Adaptive Tracking Control of Surface Vessel Using Optimized Backstepping Technique. *IEEE Trans. Cybern.* **2019**, *49*, 3420–3431. [CrossRef] [PubMed]
13. Tu Vu, V.; Pham, T.L.; Dao, P.N. Disturbance observer-based adaptive reinforcement learning for perturbed uncertain surface vessels. *ISA Trans.* **2022**, *130*, 277–292. [CrossRef] [PubMed]
14. Sutton, R.; Barto, A.G. *Reinforcement Learning*; MIT Press: Cambidge, MA, USA, 2020; p. 547.
15. Gorobtsov, A.S.; Karcov, S.K.; Pletnev, A.E.; Polyakov, Y.A. *Komp'yuternye Metody Postroeniya i Issledovaniya Matematicheskih Modelej Dinamiki Konstrukcij Avtomobilej*; Nauchno-Tekhnicheskoe Izdatel'stvo "Mashinostroenie": Moscow, Russia, 2011; p. 462. (In Russian)
16. Pogorelov, D. Differential–algebraic equations in multibody system modeling. newblock *Numerical Algorithms* **1998**, *19*, 183–194. [CrossRef]
17. Wittenburg, J. *Dynamics of Systems of Rigid Bodies*; Leitfäden der Angewandten Mathematik und Mechanik; Vieweg+Teubner Verlag: Wiesbaden, Germany, 1977.
18. Gorobtsov, A.S.; Skorikov, A.V.; Tarasov, P.S.; Markov, A.; Dianskij, A. Metod sinteza programmnogo dvizheniya robotov s uchetom zadannyh ogranichenij reakcij v svyazyah. In Proceedings of the XIII Vserosijskaia Nauchno Tekhnicheskaia Konferencia s Mezhdunarodnym Uchastiem "Robototekhnika i Iskusstvennyj Intellekt", Krasnoyarsk, Russia, 27 November 2021; pp. 199–203. (In Russian)
19. Mamedov, S.; Khusainov, R.; Gusev, S.; Klimchik, A.; Maloletov, A.; Shiriaev, A. Underactuated mechanical systems: Whether orbital stabilization is an adequate assignment for a controller design? *IFAC-PapersOnLine* **2020**, *53*, 9262–9269. [CrossRef]
20. FRUND—A System for Solving Non-Linear Dynamic Equations. Available online: http://frund.vstu.ru/ (accessed on 24 October 2022).
21. Raibert, M.H. Legged Robots. *Commun. ACM* **1986**, *29*, 499–514. [CrossRef]
22. Kim, J.Y.; Park, I.W.; Oh, J.H. Experimental realization of dynamic walking of the biped humanoid robot KHR-2 using zero moment point feedback and inertial measurement. *Adv. Robot.* **2006**, *20*, 707–736. [CrossRef]
23. Gorobtsov, A.; Andreev, A.; Markov, A.; Skorikov, A.; Tarasov, P. Features of solving the inverse dynamic method equations for the synthesis of stable walking robots controlled motion. *Inform. Autom.* **2019**, *18*, 85–122. (In Russian) [CrossRef]
24. Englsberger, J.; Werner, A.; Ott, C.; Henze, B.; Roa, M.A.; Garofalo, G.; Burger, R.; Beyer, A.; Eiberger, O.; Schmid, K.; et al. Overview of the torque-controlled humanoid robot TORO. In Proceedings of the 14th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2014, Madrid, Spain, 18–20 November 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 916–923. [CrossRef]
25. Englsberger, J.; Ott, C.; Albu-Schäffer, A. Three-dimensional bipedal walking control using Divergent Component of Motion. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 2600–2607. [CrossRef]
26. Pratt, J.E.; Carff, J.; Drakunov, S.V.; Goswami, A. Capture Point: A Step toward Humanoid Push Recovery. In Proceedings of the 2006 6th IEEE-RAS International Conference on Humanoid Robots, Genova, Italy, 4–6 December 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 200–207. [CrossRef]
27. Englsberger, J.; Koolen, T.; Bertrand, S.; Pratt, J.E.; Ott, C.; Albu-Schäffer, A. Trajectory generation for continuous leg forces during double support and heel-to-toe shift based on divergent component of motion. In Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 4022–4029. [CrossRef]
28. Khusainov, R.; Klimchik, A.; Magid, E. Swing Leg Trajectory Optimization for a Humanoid Robot Locomotion. In Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016), Lisbon, Portugal, 29–31 July 2016; Gusikhin, O., Peaucelle, D., Madani, K., Eds.; SciTePress: Setúbal, Portugal, 2016; Volume 2, pp. 130–141. [CrossRef]
29. Khusainov, R.; Shimchik, I.; Afanasyev, I.; Magid, E. Toward a Human-like Locomotion: Modelling Dynamically Stable Locomotion of an Anthropomorphic Robot in Simulink Environment. In Proceedings of the ICINCO 2015—12th International Conference on Informatics in Control, Automation and Robotics, Colmar, Alsace, France, 21–23 July 2015; Filipe, J., Madani, K., Gusikhin, O.Y., Sasiadek, J.Z., Eds.; SciTePress: Setúbal, Portugal, 2015; Volume 2, pp. 141–148. [CrossRef]
30. Khusainov, R.; Afanasyev, I.; Sabirova, L.; Magid, E. Bipedal robot locomotion modelling with virtual height inverted pendulum and preview control approaches in Simulink environment. *J. Robot. Netw. Artif. Life* **2016**, *3*, 182–187. [CrossRef]
31. Khusainov, R.; Afanasyev, I.; Magid, E. Anthropomorphic robot modelling with virtual height inverted pendulum approach in Simulink: step length and robot height influence on walking stability. In Proceedings of the ICAROB 2016—International Conference on Artificial Life and Robotics, Okinawa Convention Center, Ginowan, Japan, 29 January 2016; Volume 21, pp. 208–211. [CrossRef]
32. Liu, C.; Wang, D.; Chen, Q. Central Pattern Generator Inspired Control for Adaptive Walking of Biped Robots. *IEEE Trans. Syst. Man Cybern. Syst.* **2013**, *43*, 1206–1215. [CrossRef]
33. Glazunov, V. *Mekhanizmy Parallel'noj Struktury i ih Primenenie: Robototekhnicheskie, Tekhnologicheskie. Medicinskie, Obuchayushchie Sistemy*; Izhevskij Institut Komp'yuternyh Issledovanij: Izhevsk, Russia, 2018. (In Russian)
34. Ganiev, R.F.; Glazunov, V.A. Manipulyacionnye mekhanizmy parallel'noj struktury i ih prilozheniya v sovremennoj tekhnike. *Dokl. Akad. Nauk.* **2014**, *459*, 428. (In Russian) [CrossRef]

35. Glazunov, V.A. *Mekhanizmy Perspektivnyh Robototekhnicheskih Sistem*; Tekhnosfera: Moskva, Russia, 2020; p. 296. (In Russian)
36. Qi, Q.; Lin, W.; Guo, B.; Chen, J.; Deng, C.; Lin, G.; Sun, X.; Chen, Y. Augmented Lagrangian-Based Reinforcement Learning for Network Slicing in IIoT. *Electronics* **2022**, *11*, 3385. [CrossRef]
37. Kamikokuryo, K.; Haga, T.; Venture, G.; Hernandez, V. Adversarial Autoencoder and Multi-Armed Bandit for Dynamic Difficulty Adjustment in Immersive Virtual Reality for Rehabilitation: Application to Hand Movement. *Sensors* **2022**, *22*, 4499. [CrossRef] [PubMed]
38. Pontryagin, L.S.; Boltyanskij, V.G.; Gamkrelidze, R.V.; Mishchenko, E.F. *Matematicheskaya Teoriya Optimalnih Processov*; Fizmatgiz: Moscow, Russia, 1961; p. 391. (In Russian)
39. Bellman, R. *Dynamic Programming*, 1st ed.; Princeton University Press: Princeton, NJ, USA, 1957.
40. Kolesnikov, A.A.; Kolesnikov, A.A.; Kuz'menko, A.A. Metody AKAR i AKOR v zadachah sinteza nelinejnyh sistem upravleniya. *Mekhatronika Avtomatizaciya Upravlenie* **2016**, *17*, 657–669. (In Russian) [CrossRef]
41. Frolov, K. Umen'shenie amplitudy kolebanij rezonansnyh sistem putem upravlyaemogo izmeneniya parametrov. *Mashinovedenie* **1965**, *3*, 38–42. (In Russian)
42. Dmitriev, A.A. *Teoriya i Raschet Nelinejnyh Sistem Podressorivaniya Gusenichnyh Mashin*; Mashinostroenie: Moscow, Russia, 1976. (In Russian)
43. Gorobtsov, A. Issledovanie vozmozhnostej sistemy vibrozashchity so stupenchato izmenyayushchimisya parametrami. In Proceedings of the IV Vsesoyuz. Simpozium Vliyanie Vibracii na Organizm Cheloveka i Problemy Vibrozashchity, Moscow, Russia, 18–20 July 1982; pp. 74–75. (In Russian)
44. Karnopp, D.; Rosenberg, R. *Analysis and Simulation of Multiport Systems: The Bond Graph Approach to Physical System Dynamics*; MIT Press: Cambidge, MA, USA, 1968; p. 220.
45. Vukobratovic, M.; Stokić, D.; Kirćanski, N. *Non-Adaptive and Adaptive Control of Manipulation Robots*; Communications and Control Engineering Series; Springer: Berlin, Germany, 1985.