Modeling How To Catch Flying Objects: Optimality Vs. Heuristics

Modellierung des Fangens fliegender Objekte: Optimalität vs. Heuristiken.
Bachelor-Thesis von Annemarie Mattmann
Tag der Einreichung:

1. Gutachten: Prof. Dr. Jan Peters
2. Gutachten: Prof. Dr. Gerhard Neumann

TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich Informatik
Fachgebiet
Intelligente Autonome Systeme

Modeling How To Catch Flying Objects: Optimality Vs. Heuristics
Modellierung des Fangens fliegender Objekte: Optimalität vs. Heuristiken.

Vorgelegte Bachelor-Thesis von Annemarie Mattmann

1. Gutachten: Prof. Dr. Jan Peters
2. Gutachten: Prof. Dr. Gerhard Neumann

Tag der Einreichung:

Erklärung zur Bachelor-Thesis

Hiermit versichere ich, die vorliegende Bachelor-Thesis ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 10th October 2014

_____

(Annemarie Mattmann)

Contents

List of Figures

Abstract

The question how humans manage to catch flying objects is one which has long been discussed in research but still lacks a sufficient answer. It is not a trivial one because noise influences the flying object thereby changing its trajectory and both uncertainty and latency influence the human's observations. Moreover, the task is time-critical especially for sports like table tennis where prediction and open loop behaviour is required as opposed to baseball where closed loop behaviour suffices. Though whether table tennis or baseball the goal of reaching the impact position in time to catch the ball remains the same.

Heuristics were defined which describe both open loop and closed loop behaviour relying on the prediction of the ball's trajectory or on online updates regarding the ball's movement. Up to now no heuristic or common framework exists which combines open loop and closed loop behaviour and the latter is deemed non-optimal while the first is not generally acknowledged.

This thesis tries to close the gap between open loop and closed loop heuristics through the consideration of noise and latency in the model. By approaching the question as an optimal control problem an optimal common framework shall be found. This shall combine all human catching techniques and include an optimal policy for human ball catching that is assumed to follow heuristics.

1 Introduction

How do humans run to catch flying objects? The basic scenario behind this question includes a human and a flying object, e.g. a ball. The ball is accelerated by some force and travels through the air until it touches the ground or is caught by the human. The human runs to the impact point and is assumed to catch the ball if he gets to that position before or at the moment the ball arrives (i.e. hand movements involved to catch the ball are not considered).

The question is how the human knows where to run because neither the environment nor the human are perfect and without limitations. First of all, the trajectory of the ball might change due to noise like air resistance or simply due to different initial conditions like a (slightly) altered throwing angle. Secondly, the human's observations of the ball are not perfectly precise including some noise by themselves and thirdly, the human's reactions are delayed by the latency of the human's perceptual and motor system.

On the one hand, these difficulties seem to call for closed loop policies that rely on the collection of many observations in order to estimate the ball's trajectory. On the other hand, humans manage to catch balls open loop when very fast movements are required to do so. Typically, it is believed that humans utilize different heuristics to solve the problem of ball catching. In this thesis, the human behaviour will be investigated from an optimality point of view and analysed for an underlying reward function that explains the behaviour.

The question how humans run to catch balls is an interesting one because most people cannot tell how they manage the task and observing them yields puzzling insights. For example, it seems rather peculiar, especially for such a time-critical task as catching a ball, that players slow down when they do not have to run as far to arrive at the impact position at exactly the same time the ball does instead of running faster and waiting for it (McLeod and Dienes, 1996). Players behave intuitively but observers might have different ideas of how the task should be done and accuse the players for behaving wrongly. A coach will probably blame them for being lazy or taking a risk by running too slow or into the wrong direction instead of running at maximum speed on a straight trajectory.

The coach is asking for what he deems is optimal behaviour but it implies that the players know where the ball will touch the ground immediately after it is thrown. However, this knowledge involves solving complex differential equations which need to regard not only the exact initial position and speed of the ball and the angles of the throw but also air resistance, spin, wind and every other factor which might influence the ball's trajectory (Dawkins, 2006, Adair, 1990, McBeath, Nathan, Bahill, and Baldwin, 2008). Of course all of these values would need to be determined first and very precisely because small differences in the initial values can lead to large errors in the calculated impact position. The theory behind this is called *Trajectory Prediction* (TP) and states that players predict the impact position of the ball (Saxberg, 1987).

According to Gigerenzer et al. (see Marewski, Gaissmaier, and Gigerenzer, 2009), this strategy would be (time-)optimal but humans are most likely not capable of succeeding in the task. This view is supported by the fact that players fail when they need to predict impact positions or recognize the correct ball trajectories (Shaffer and McBeath, 2005). These facts suggest that they are not able to catch a ball utilizing TP by pre-calculating the impact position and running there as fast as possible because of the noise in the system and observations which lead to inaccurate predictions. Instead, Gigerenzer et al. state that whenever uncertainty or incomplete knowledge pose a significant problem for solving a task, humans achieve (more) accurate results by utilizing reactive heuristics.

Regarding the problem at hand, the uncertainty is a result of the different initial conditions (changing throw angles etc.) and the noise induced into the system by both the ball (air resistance etc.) and the player (measurement errors regarding the ball's and probably his own position). The uncertainty is higher the longer the ball is in the air, because it is constantly influenced by noise. Thus, for sports like baseball according to Gigerenzer et al.'s thesis a player should apply a reactive heuristic, because it is less error-prone than Trajectory Prediction.

It is indeed speculated that a player watches the ball while he runs and chooses his running speed and trajectory through online updates regarding the changes of the ball's position. These updates enable the player to couple his movements with the movement of the ball to maintain a "collision course" which eliminates the necessity of computations (Fajen and Warren, 2007, Chohan, Verheul, Kampen, Wind, and Savelsbergh, 2008, Fink, Foo, and Warren, 2009, Shaffer, Krauchunas, Eddy, and McBeath, 2004, Marewski, Gaissmaier, and Gigerenzer, 2009). Three such reactive heuristics exist, namely *Optical Acceleration Cancellation*, *Constant Bearing Angle* and *Linear Optical Trajectory*.

According to *Optical Acceleration Cancellation* (OAC) to catch a ball the human maintains a running speed such that the rate of change of the tangent of the ball's elevation angle remains constant (Chapman, 1968), i.e. the distance between the ball and the human on the x-axis changes proportionally to the height of the ball above ground. In other words by

perceiving changes in the ball's velocity[1] the human tries to run such that he keeps the vertical optical acceleration of the ball constant (see Figure 1.1a and 1.1b).

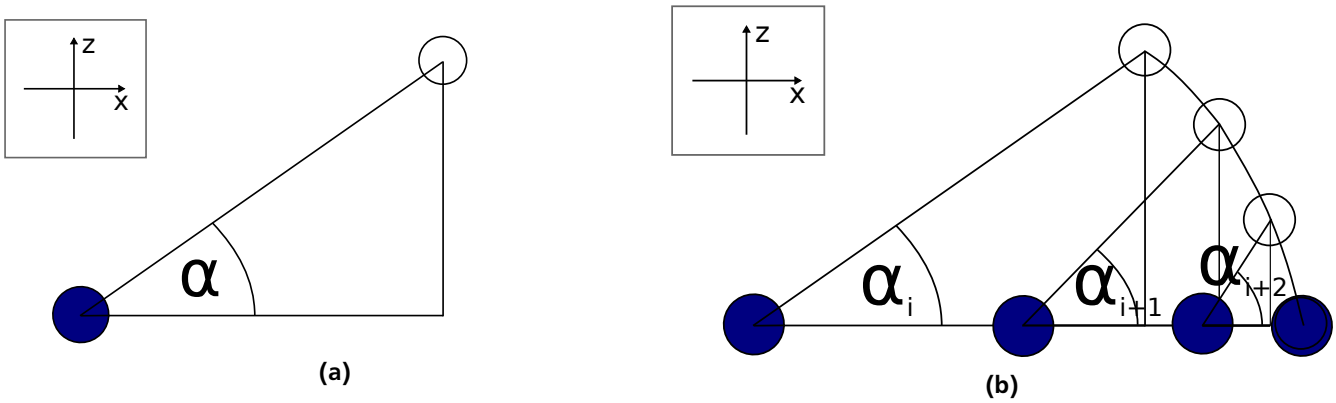


Figure 1.1.: (a) The concept of OAC showing the elevation angle α, which is the vertical angle between the human (filled out circle), the ball and the ground. Following the OAC theory, the human tries to keep the rate of change of $\tan(\alpha)$ constant (i.e. keep the ball at constant speed). (b) An example trajectory for OAC showing the last four time steps of which the last one shows a catch. The human slows down when he comes near to the impact position (as can be seen by the ever smaller distance he travels in each time step), keeping the rate of change of the tangent of the elevation angles α_i constant to cancel out vertical optical acceleration of the ball.

However, OAC alone only applies to ball trajectories which align with the human. An additional condition is needed to cover lateral movement for balls heading to the side of the human (Chapman, 1968). The *Constant Bearing Angle* (CBA) theory, which was proposed by Chapman along with the OAC theory in 1968, states that to be at the impact position of the ball at the right time regarding the horizontal plane the human needs to adapt his running speed such as to keep the bearing angle of the ball constant (i.e. keep the ball due a certain compass point) as illustrated in Figure 1.2a and 1.2b.

CBA has long been applied for collision avoidance in sailing and flight (Fajen and Warren, 2007, Chohan, Verheul, Kampen, Wind, and Savelsbergh, 2008, Diaz, Phillips, and Fajen, 2009, Marewski, Gaissmaier, and Gigerenzer, 2009). When another ship or plane approaches, the sailor or pilot keeps his course and checks whether the bearing angle remains the same. If it does an evasive manoeuvre is performed immediately to avoid collision. This strategy can be applied conversely for tasks such as tracking and ball catching where collision is desired (Chapman, 1968).

CBA does not deliver a unique solution by itself (see Figure 1.2b). It even includes the straight line trajectory which is not generally observed in human ball catching behaviour. Instead, it is an addition to OAC to cover lateral movement where OAC dictates the running speed of the human and thus the trajectory resulting from CBA.

Another approach for ball catching is provided by the *Linear Optical Trajectory* (LOT) theory, where humans try to keep the optical trajectory projection angle γ (see Figure 1.3a and 1.3b) constant (McBeath, Shaffer, and Kaiser, 1995, McLeod, Reed, and Dienes, 2001). The angle γ is the observed angle of the ball movement relative to the background horizon and lies between the background horizon, the initial ball position and the projection of the ball onto the background image as seen by the human's point of view. A human following the LOT strategy selects a running path which keeps the angle of ball movement (γ) constant on the 2D-plane projection of the ball against the background horizon and, thus, maintains a linear optical ball trajectory relative to the initial ball position and the background horizon. The angle γ will be held constant if $\tan(\alpha)$ and $\tan(\beta)$ change proportionally to each other over time with $\tan(\gamma) = \frac{\tan(\alpha)}{\tan(\beta)}$ where α is the elevation angle and β is the angle between the initial ball position, the human and the ball on the horizontal plane (see Figure 1.3b).

More recently, OAC is referred to as an addition for the LOT theory when regarding the vertical movement of balls (e.g. Shaffer and McBeath, 2002, Shaffer et al., 2004). Thus OAC with CBA, OAC with LOT and LOT alone are the catching strategies to be considered in a three-dimensional space.

An important aspect to notice is that none of the theories directly enforce a certain direction in which the human is facing as long as the ball is somewhere in his field of view.

Further investigation of the theories supports the applicability of both OAC and LOT for catching ground balls, i.e. balls rolled across the ground instead of thrown in the air (Sugar, McBeath, and Wang, 2006b), and for dogs catching frisbees (Shaffer et al., 2004), which suggests that animals utilize the same techniques as humans and that the theories may be generalized for all flying and non-flying objects even if they travel on complex trajectories such as frisbees. Both theories were also maintained when trying to catch uncatchable balls (Shaffer and McBeath, 2002) and applied for mobile robot

[1] Since humans have trouble with the perception of change in the acceleration (Fink, Foo, and Warren, 2009, Calderone and Kaiser, 1989, Schmerler, 1976).

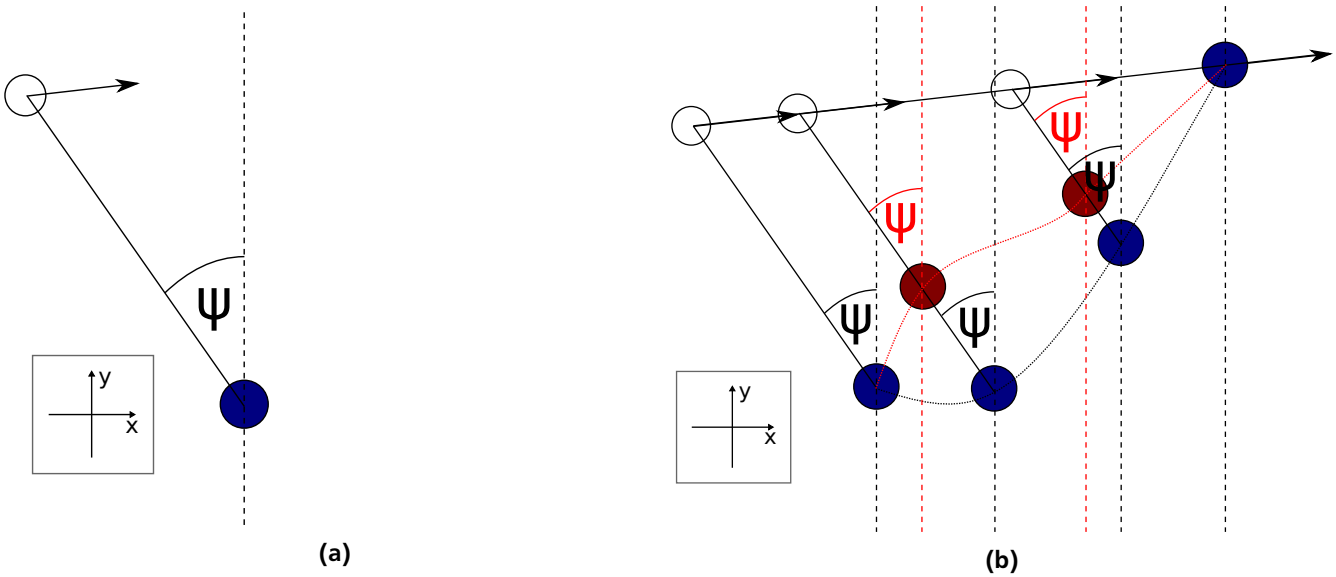


Figure 1.2.: (a) The concept of CBA showing the bearing or azimuth angle ψ, which is the horizontal angle between the human (filled out circle), the ball and a predefined reference line (the striped line from north to south). Following the CBA theory the human tries to keep ψ constant. (b) Two (exaggerated) example trajectories for CBA showing a constant ψ which results in a catch as represented in step 4 (note that the steps are not uniform distance time steps). Step 2 of the red trajectory illustrates that the human is not required to run on a straight line to the impact position (and will most likely refrain from doing so) as long as he keeps the bearing angle constant. The variations in the human's position can be compensated by changes in running speed so apart from the two solutions shown here, many different trajectories can result from keeping the bearing angle constant.

design (Sugar, McBeath, Suluh, and Mundhra, 2006a, Sugar and McBeath, 2001b,a). CBA on the other hand has been deployed by sailors and pilots for collision detection (Fajen and Warren, 2007, Chohan, Verheul, Kampen, Wind, and Savelsbergh, 2008) as well as by humans navigating around obstacles and tracking a target (Fajen and Warren, 2007, Chapman, 1968).

However, when put to the test in a virtual environment, where the ball's trajectory was altered during mid-flight thus creating a trajectory, which would not be possible in a world that underlies our laws of physics, LOT breaks down (Fink, Foo, and Warren, 2009). Though the humans are still able to catch the balls, they do not utilize the LOT theory nor are they able to calculate the ball's trajectory in advance. Instead they seem to draw upon OAC and CBA for catching.

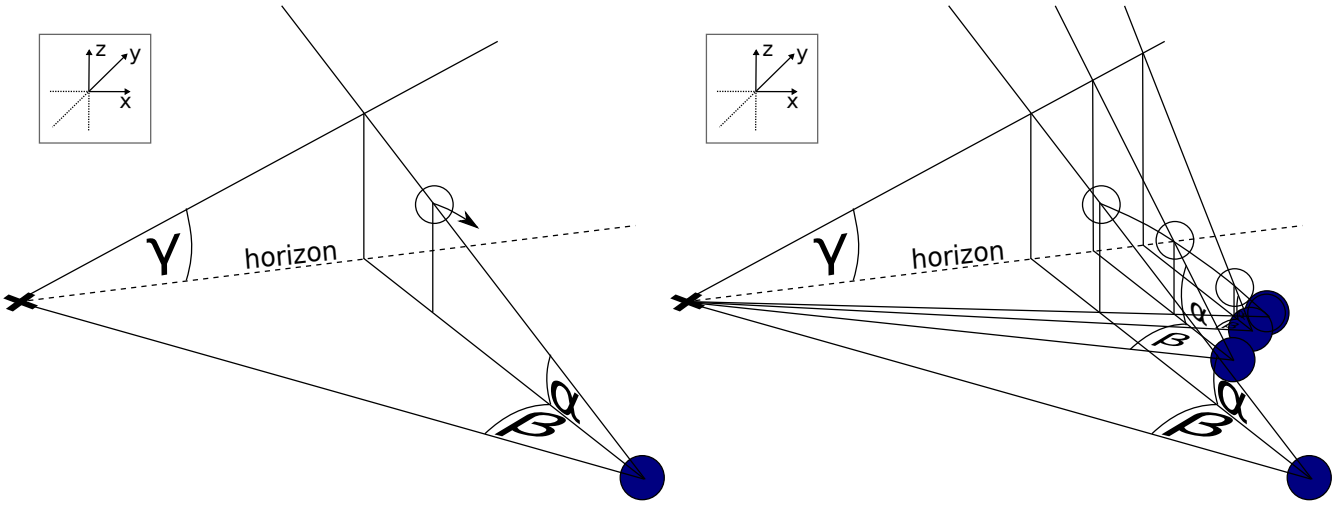
Still, the reactive heuristics seem to work. However, there are scenarios such as table tennis where the ball is too fast for the player to react in time due to the latency of the human motor system and the short travel time of the ball. There the reactive heuristics fail and a prediction of the ball's trajectory is required. Such prediction also seems appropriate because of the lower uncertainties (i.e. there is no wind, little air resistance and the ball is so near that the human will have a good estimate of its position).

Furthermore, scenarios where the ball is flying at high speed (e.g. tennis) need to be considered. Players might be forced to intermediately stop their observation of the ball to run faster while relying on a prediction of the impact area only until the ball returns into their field of view. Hence, Trajectory Prediction also needs to play a role if the amount of uncertainty is small and high latency requires prediction as the only remaining choice when reactive heuristics fail due to high latencies.

In this thesis, a different point of view is taken than is discussed in previous research. The assumption is that the human optimizes some reward function, i.e., he tries to act optimally. However, due to his constraints (latency and uncertainty) this behaviour does not look optimal and it reproduces the behaviour of heuristics. Hence, the main assumption of this thesis is that given the constrains of uncertainty and latency the human behaviour (of following a heuristic) is optimal.

It is further assumed that this behaviour includes both predictive and reactive aspects which combine to a common framework for all human ball catching techniques and which can be modelled by an optimality criterion. This framework is assumed to apply to all scenarios described above when only noise and delay are changed to reflect uncertainties and latencies in different catching situations like baseball, tennis and table tennis. This idea is derived from the facts listed above which concern the capability of humans being able to play both baseball and table tennis though they seem to have very different requirements.

In order to investigate these hypotheses, a model for human ball catching is defined, which includes both uncertainties and latencies. For an initial analysis, the model is simplified to a two-dimensional problem formulation neglecting



(a) (b)

Figure 1.3.: (a) The concept of LOT showing the elevation angle α, the angle β between the initial ball position (marked with an **x**), the human and the ball on the horizontal plane and the optical trajectory projection angle γ between the background horizon, the initial ball position and the projection of the ball onto the background image as seen by the human. The angle γ should be constant while $\tan(\alpha)$ and $\tan(\beta)$ should change proportionally to each other. (b) An example trajectory for LOT where the human curves in to the impact position and catches the ball. The angle γ shown against the horizon remains constant while $\tan(\alpha)$ and $\tan(\beta)$ increase proportionally to each other leading to the trajectory curvature that is typical for LOT (McBeath, Shaffer, and Kaiser, 1995).

latencies as well as the agent's (i.e. the human's) field of view. In order to deal with the uncertainties in the observations the Kalman Filter (Kalman, 1960) is utilized. To obtain an optimal policy Stochastic Optimal Control (Movellan, 2009) is investigated and replaced by CMA-ES[2] (Hansen and Ostermeier, 1996) in a second, more complex three-dimensional model, which is defined to include latencies and a field of view.

The resulting optimal policy is evaluated in relation to the heuristics mentioned above to show whether the observed behaviour of the human can be reproduced with optimal control. Additionally, the policy is tested with different settings for noise and delay to find universal strategies for the catching scenarios (of e.g. baseball, tennis and table tennis).

The next section includes the description of the models. Section 3 describes the state estimation and learning methods utilized for the optimal control problem. Section 4 includes the evaluation and discussion of the results and Section 5 will give a short recap and proposals for future work.

[2] Covariance Matrix Adaptation Evolution Strategy

2 Computational Models for Ball-Catching

In the following the two models for human ball catching will be described. The ball and the human are modelled both times as a dynamical system

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t),$$

with (partial) observations

$$\mathbf{y}_t = g(\mathbf{x}_t),$$

from which the optimal policy will be derived to guide the human's movement through the state space for each time step.

The first model does not suffice to meet all requirements given in the introduction but rather serves as a first insight into the problem. The second more complex model is defined to fully cover all important aspects of human ball catching as mentioned above and to allow the analysis of the posed hypotheses.

2.1 Simplified Ball Catching Model

The first model is rather simple with a two-dimensional state space where the x-axis represents the horizontal and the y-axis represents the vertical direction as shown in Figure 2.1. The state space is defined as the position and velocity of the human and the ball. A catch occurs when the human is at the same place as the ball when the latter touches the ground (crosses the x-axis).

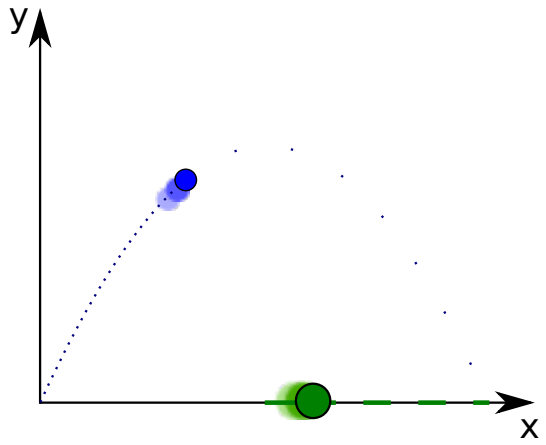


Figure 2.1.: A sketch of the model in an example scenario. The ball is in the air shown as a blue circle and the human is shown as a larger green circle on the ground (i.e. on the x-axis). The human is already in motion running to catch the ball at its impact position.

2.1.1 Model Description

To fully describe the simplified two-dimensional model of a human catching a ball, a reward function $r(\mathbf{x})$, observations $\mathbf{y}$, motor commands $\mathbf{u}$ and a system state $\mathbf{x}$ modelled as the linear system

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{g} + \boldsymbol{\omega}_t,$$

with the system error $\boldsymbol{\omega} \sim \mathcal{N}(0, \mathbf{R})$, the system matrices $\mathbf{A}$ and $\mathbf{B}$ and the system vector $\mathbf{g}$ need to be defined. A linear system is chosen because it is required by the Kalman Filter and while it is rather simple it also suffices for the simplified problem definition and can be optimized efficiently with Stochastic Optimal Control.

The state vector $\mathbf{x}$ is given by

$$\mathbf{x} = \begin{bmatrix} x_B & y_B & \dot{x}_B & \dot{y}_B & x_H & \dot{x}_H \end{bmatrix},$$

where x_B and y_B are the x- and y-position of the ball, $\dot{x}_B$ and $\dot{y}_B$ are the velocities of the ball and x_H and $\dot{x}_H$ are the position and velocity of the human.

Modelling the Human

The human is regarded as a point mass (rigid body). He can only move along the x-axis of the two-dimensional state space. The initial velocity of the human is zero (i.e. he stands still) and the human acceleration is given by

$$\ddot{x}_H = u + \omega_{x_H},$$

with Gaussian noise ω_{x_H}, which is part of the system error $\boldsymbol{\omega}$.

The motor commands are calculated by a PD-controller (proportional-derivative controller)

$$\mathbf{u}_t = \mathbf{K}_t \mathbf{x}_t + \mathbf{k}_t,$$

which is the optimal policy (with optimal gains $\mathbf{K}_t$ and $\mathbf{k}_t$) gained by the optimal control algorithm and allows for stable trajectory tracking (Kawarnura, Miyazaki, and Arimototest, 1988, Tomei, 1991). In the simplified model, the motor commands are unconstrained and the human is assumed to see the ball independent of its position, i.e. the human's field of view is neglected.

Modelling the Ball

The ball is regarded as a point mass and can move in x- and y-direction until its y-value becomes zero or negative, which is defined as an impact on the ground and the simulation is stopped.

Air resistance, spin and other factors which could influence the ball's trajectory are not taken into account directly but are included as Gaussian noise. However, gravity is taken into account as negative vertical acceleration of the ball. The accelerations are given by

$$\begin{aligned} \ddot{x}_B &= \omega_{x_B}, \\ \ddot{y}_B &= -g + \omega_{y_B}, \end{aligned}$$

with Gaussian noise ω_{x_B}, ω_{y_B}, which is part of the system error $\boldsymbol{\omega}$ and the gravitational acceleration constant $g \approx 9.81\,\mathrm{m\,s^{-2}}$.

Calculating the System Matrices

The matrices $\mathbf{A}$, $\mathbf{B}$ and the vector $\mathbf{g}$ are set according to the semi-implicit or symplectic Euler integration method (Hairer, Lubich, and Wanner, 2003)

$$\begin{aligned} \dot{x}_{t+1} &= \dot{x}_t + \Delta t \ddot{x}_t, \\ x_{t+1} &= x_t + \Delta t \dot{x}_{t+1}, \end{aligned}$$

which is more stable than the standard Euler method while equally complex and costly. The values for the matrices can be looked up in Appendix A.

Modelling the Human Observation

The human is unable to perceive the true system state $\mathbf{x}$ but he observes the position of the ball in every time step

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \boldsymbol{\nu}_t,$$

where $\mathbf{y}$ is the observation, $\boldsymbol{\nu} \sim \mathcal{N}(0, \mathbf{Q})$ is the measurement error and $\mathbf{C}$ is the matrix to extract the relevant entries from the hidden state vector $\mathbf{x}$ (the matrix values are given in Appendix A). In this case, only the ball's position and the human position are extracted, the ball's velocity and acceleration as well as the human velocity are assumed to be observed perfectly.

Since latency is not considered for the simplified model the observations gained at each time step are undelayed.

Modelling the Reward

Desired behaviour as it is considered for the problem at hand is that the human catches the ball (which is his only goal). A successful catch is defined by $x_{B_T} \approx x_{H_T}$ (the approximate equal is referring to an error of at most 20 cm), i.e. the human being at the same place as the ball at the last time step T, which is given by the time when either the ball touches the ground or it is caught by the human.

To gain the desired behaviour, a reward function needs to be found which makes this behaviour optimal. In order to achieve this, large distances from the impact position are punished at the time the ball touches the ground. Additionally, high motor commands are punished for every time step to ensure smooth trajectories.

A quadratic reward function (see Movellan, 2009)

$$r(\mathbf{x}, \mathbf{u}) = -\mathbf{x}^T \mathbf{G}\mathbf{x} - \mathbf{u}^T \mathbf{H}\mathbf{u}$$

is used, where **G** and **H** represent weights. **G** is the reward matrix for the state and **H** is the reward matrix for the controls. Together, they define a trade-off between reaching a certain state and requiring a minimum amount of motor commands.

Since a reward is given only for catching the ball (with an error of at most 20 cm), **G** is always zero but for the last time step at which either an impact or a catch occurs and the simulation is stopped. The punishment is given by the negative quadratic distance between human and ball $-(x_{B_T} - x_{H_T})^2$ for the last time step T resulting in the matrix form of $\mathbf{G}_T$ given in Appendix A.

H is tuned for optimal behaviour to achieve a smooth trajectory. It's value can also be looked up in A.

Resulting Model

Summing up all the above definitions yields the final simplified model that describes the optimal control problem (for details check the above explanations).

There is the dynamical system described as

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{g} + \boldsymbol{\omega}_t,$$

with the state

$$\mathbf{x} = \begin{bmatrix} x_B & y_B & \dot{x}_B & \dot{y}_B & x_H & \dot{x}_H \end{bmatrix}$$

and the motor command

$$\mathbf{u} = \begin{bmatrix} \ddot{x}_H \end{bmatrix}.$$

The observation at time step t is computed by

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \boldsymbol{\nu}_t.$$

The reward function is defined as

$$r(\mathbf{x}, \mathbf{u}) = -\mathbf{x}^T \mathbf{G}\mathbf{x} - \mathbf{u}^T \mathbf{H}\mathbf{u}.$$

Together, these definitions are assumed to form the minimal model for the simulation of human ball catching, neglecting latency and field of view.

2.2 Complex Model With Latencies and Field of View

The model described in the last section already includes human uncertainty and permits the simulation of catching a flying object in a two-dimensional state space but it does not fully meet the requirements. To reproduce the behaviour of the heuristics described in the introduction a three-dimensional state space is required. Furthermore, to reproduce the different circumstances regarding different ball catching scenarios (baseball, tennis, table tennis) the human's latency also needs to be regarded. Additionally, to reflect the constraints of reality a field of view and limits for the motor commands should be considered. Thus in the following a more elaborate model is established.

The state space of the new model is three-dimensional with the x- and y-axis representing the horizontal and the z-axis representing the vertical directions as shown in Figure 2.2. The state space is defined as the position of the human and the position and velocity of the ball. A catch occurs when the human is at the same place as the ball when the latter touches the ground (crosses the xy-plane).

2.2.1 Model Description

To fully describe the three-dimensional model of a human catching a ball, a reward function $r(\mathbf{x})$, observations **y**, (limited) motor commands **u** and a system state **x** modelled as the non-linear system

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \mathbf{g} + \boldsymbol{\omega}_t,$$

with the system error $\boldsymbol{\omega} \sim \mathcal{N}(0, \mathbf{R})$, need to be defined. The non-linear system is required because the linear system from the simplified model is not sufficient to describe the influence of the viewing direction and field of view on the motor commands of the human, thus the part of the system describing the human dynamics must be non-linear with a time dependent $\mathbf{B}_t$ matrix.

The state vector **x** is given by

$$\mathbf{x} = \begin{bmatrix} x_H, & y_H, & \phi_H, & x_B, & y_B, & z_B, & \dot{x}_B, & \dot{y}_B, & \dot{z}_B \end{bmatrix},$$

where x_H and y_H are the x- and y-position of the human, ϕ_H is the viewing direction of the human, x_B, y_B and z_B are the ball's x-, y- and z-position and $\dot{x}_B$, $\dot{y}_B$ and $\dot{z}_B$ are the velocities of the ball.

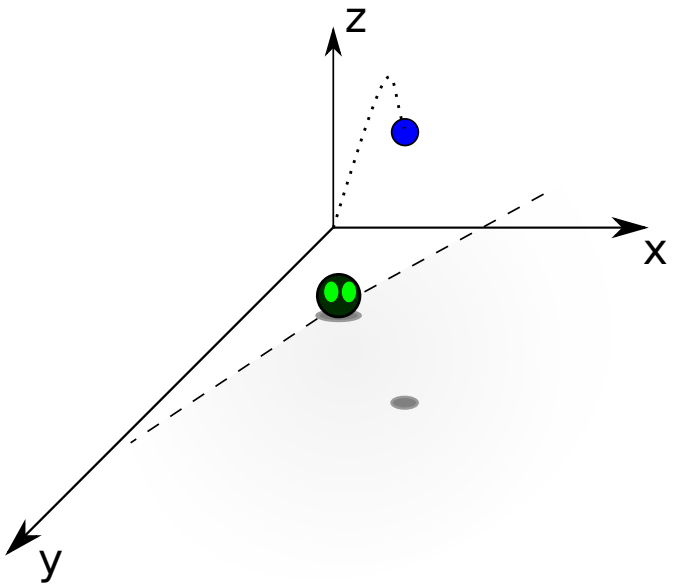


Figure 2.2.: A sketch of the new model in an example scenario. The ball, which is shown by a small blue sphere, is in the air as can be told by its shadow. The human is depicted as the larger green sphere with eyes and is positioned on the ground (i.e. the xy-plane) looking into the direction of the ball. The field of view is depicted by the dashed lines originating from the human and a slight shading in between.

Modelling the Human

The human is regarded as a point mass (rigid body). He can only move in the xy-plane of the three-dimensional state space. The human's view is bounded to a 175° field of view in front of him to reflect natural limitations of sight (see Figure 2.2). In this field of view the human is capable of seeing the ball independent of its position particularly regarding the z-direction, which is unbounded. Everything outside the field of view cannot be observed resulting in a higher uncertainty since the human has to rely on predictions of the ball's position. The viewing direction ϕ_H marks the middle of the field of view (see Figure 2.3).

The human's acceleration is disregarded, instead, the motor commands are given by the human's velocity, which is expressed as both translational u_T and rotational u_R movement speed (see Figure 2.3), where the x- and y-part depend on the viewing direction[1], thus,

$$\begin{aligned}\dot{x}_H &= u_T \cos\phi_H + \omega_{x_H},\\ \dot{y}_H &= u_T \sin\phi_H + \omega_{y_H},\\ \dot{\phi}_H &= u_R + \omega_{\phi_H},\end{aligned}$$

with Gaussian noise $\omega_{x_H}, \omega_{y_H}$ and ω_{ϕ_H}, which is part of the system error $\boldsymbol{\omega}$. The motor commands

$$\mathbf{u} = \left[u_T, u_R\right]$$

are again calculated by the PD-controller

$$\mathbf{u}_t = \mathbf{K}_t \mathbf{x}_t + \mathbf{k}_t,$$

though this time the gains $\mathbf{K}_t$ and $\mathbf{k}_t$ are learned with CMA-ES because the system required for the complex model is non-linear and Stochastic Optimal Control requires a linear system.

To reflect a human's speed maximum for running forwards, backwards and turning around the motor commands are limited to a range of possible values given by the intervals $u_T \in [-2.5, 5]$ (in meter/second) and $u_R \in [-2\pi, 2\pi]$ (in rad/second). Backwards running is modelled by a negative velocity, forward running by a positive one. Turning clockwise is modelled by a negative angular velocity and turning anti-clockwise by a positive one.

[1] More precisely, translational movements will move the human forwards or backwards on a straight line into the direction given by the viewing direction and rotational movements will turn the human around his own axis (with no additional forwards or backwards movement) with the viewing direction is a starting offset. Translational and rotational velocity can be combined to describe curves in the trajectory.

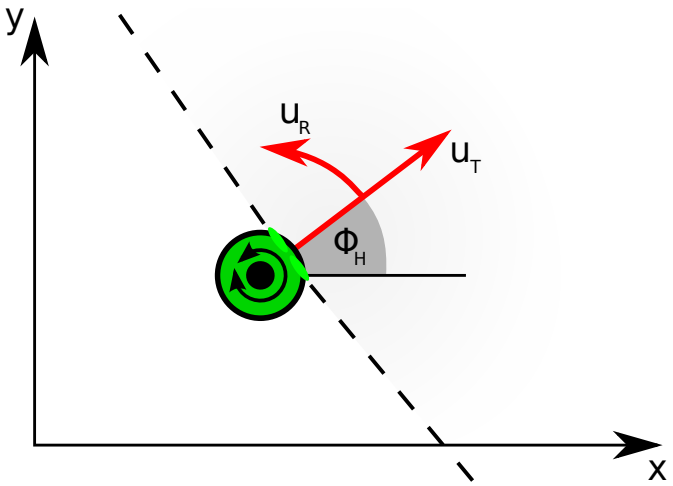


Figure 2.3.: The human is seen from above on the xy-plane. The direction of the motor commands are shown as red arrows. The translational motor command u_T will result in forward or backward movement of the human and the rotational command u_R will result in clockwise or counter-clockwise rotational movement. The center of rotation is the human point mass as sketched by the black dot and arrows. The viewing direction ϕ_H is the angle between the x-axis and the viewing direction of the human. The field of view is depicted by the dashed lines originating from the human and a slight shading in between.

Modelling the Ball

The ball is regarded as a point mass and can move in x-, y- and z-direction until its z-value becomes zero or negative, which is defined as an impact on the ground and the simulation is stopped.

Air resistance, spin and other factors which could influence the ball's trajectory are not taken into account directly but are included as Gaussian noise. However, gravity is taken into account as negative vertical acceleration of the ball. The ball's acceleration components are given by

$$\begin{aligned} \ddot{x}_B &= \omega_{x_B}, \\ \ddot{y}_B &= \omega_{y_B}, \\ \ddot{z}_B &= -g + \omega_{z_B}, \end{aligned}$$

with Gaussian noise $\omega_{x_B}, \omega_{y_B}$ and ω_{z_B}, which is part of the system error $\boldsymbol{\omega}$ and the gravitational acceleration constant $g \approx 9.81\,\mathrm{m\,s^{-2}}$.

Calculating the System Matrices

The matrices **A** and **B** and the vector **g** are again set according to the symplectic Euler integration method and can be inspected in Appendix B.

Modelling the Human Observation

The human is unable to perceive the true system state **x** but he observes the position of the ball whenever it is in his field of view of 175° total (assuming the human cannot turn his head), i.e. $\phi_{HB} \in [-1.527, 1.527]$ (where 1.527 radian corresponds to 87.5 degrees) with

$$\phi_{HB} = \operatorname{atan2}(y_b - y_H, x_B - x_H) - \phi_H,$$

which denotes the angle between the viewing direction of the human and the ball relative to the human (see Figure 2.4).

Additionally, there is a latency to the human's reactions which is expressed by a delay in the human's observation of the ball (in particular, this implies that for the first few time steps the human only has the prediction based on the mean of the initial values since every observation made is delayed). Thus, if the ball is in the human's field of view the observation is gained by

$$\mathbf{y}_{t+\tau} = \mathbf{C}\mathbf{x}_t + \boldsymbol{\nu}_t,$$

where τ is the delay in number of time step given by the human latency (e.g. a latency of 180 ms results in a delay of $\tau = 9$ time steps for a step width of $\Delta t = 0.02$), $\boldsymbol{\nu} \sim \mathcal{N}(0, \mathbf{Q})$ is the measurement error and **C** is the matrix to extract the relevant entries from the hidden state vector **x** (the matrix values are given in Appendix B). In this case only the ball's position are extracted, the ball's velocity and human's position and direction are assumed to be observed perfectly.

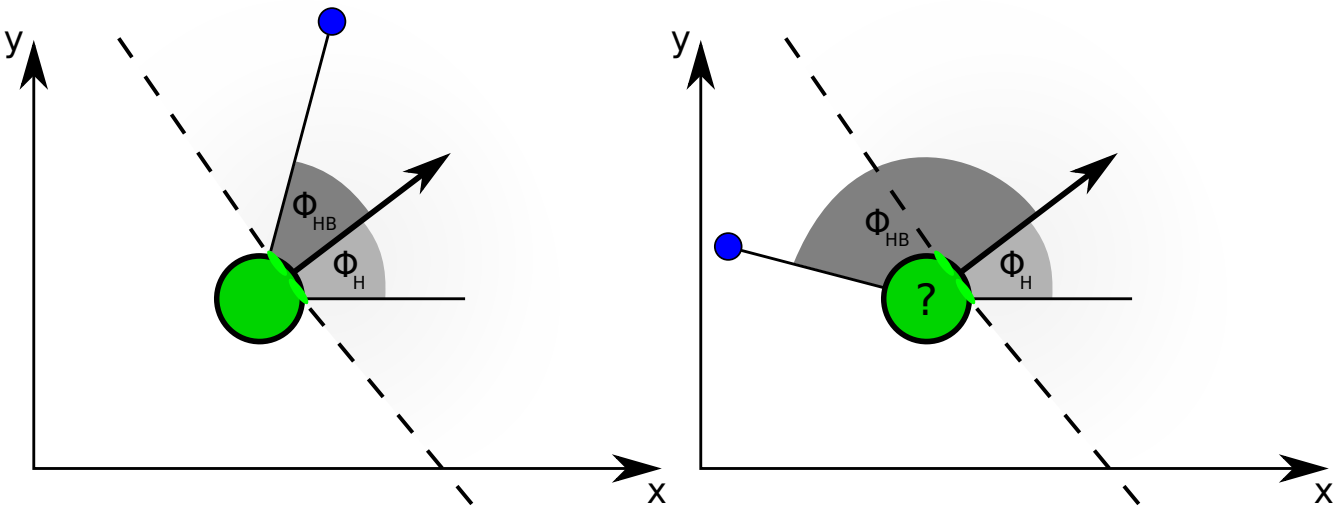


Figure 2.4.: The human is seen from above on the xy-plane. The angle ϕ_{HB} between human and ball is the angle between the viewing direction of the human and the ball relative to the human. The viewing direction ϕ_H is the angle between the x-axis and the viewing direction of the human. The field of view is depicted by the dashed lines originating from the human and a slight shading in between. In the left picture the ball is inside the field of view of the human so the human makes an observation. In the right picture the human makes no observation because the ball is outside of his field of view and he is uncertain about its position.

Modelling the Reward

The desired behaviour is that the human catches the ball and that doing so is his only goal. A successful catch is defined by $x_{B_T} \approx x_{H_T}$ and $y_{B_T} \approx y_{H_T}$ (the approximate equal is referring to a distance error of at most 20 cm), i.e. the human being at the same place as the ball at the last time step T, which is given by the time when either the ball touches the ground or it is caught by the human.

To achieve the desired behaviour, a new reward function needs to be found which makes this behaviour optimal. In order to achieve this, large distances from the impact position are punished at time step T with the negative quadratic distance between ball and human

$$r(\mathbf{x}_T) = -(x_{B_T} - x_{H_T})^2 - (y_{B_T} - y_{H_T})^2.$$

Resulting Model

Summing up all the above definitions yields the final model that describes the optimal control problem (for details check the above explanations).

There is the dynamical system described as

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \mathbf{g} + \boldsymbol{\omega}_t,$$

with the state

$$\mathbf{x} = \begin{bmatrix} x_H, & y_H, & \phi_H, & x_B, & y_B, & z_B, & \dot{x}_B, & \dot{y}_B, & \dot{z}_B \end{bmatrix}$$

and the motor commands

$$\mathbf{u} = \begin{bmatrix} u_T, u_R \end{bmatrix}.$$

The observation at time step t is computed by

$$\mathbf{y}_{t+\tau} = \mathbf{C}\mathbf{x}_t + \boldsymbol{\nu}_t.$$

The reward function is defined as

$$r(\mathbf{x}) = \begin{cases} 0, & \text{if } z_B > 0 \\ -((x_B - x_H)^2 + (y_B - y_H)^2), & \text{else.} \end{cases}$$

Together, these definitions are assumed to form the minimal model for the simulation of human ball catching.

3 Methods

In the following, the optimal state estimator for both the simplified and the complex model is introduced along with the approaches made to learn the respective optimal control policy.

3.1 Optimal State Prediction

Since the human is unable to perceive the true state $\mathbf{x}$ directly (due to his imperfect perceptual ability), he has to rely on his noisy observations and reason about the true configuration of the ball. To do so, a belief (Gaussian distribution over the states) needs to be calculated for the current configuration of the ball given by a mean and a variance, which need to be computed iteratively. The belief captures the human's knowledge on the configuration but also his uncertainty.

The Kalman Filter (Kalman, 1960, Welch and Bishop, 2006, Thrun, Burgard, and Fox, 2005) is introduced as an estimator of the true state which allows the human to predict the ball's position as it compensates for the noise that the measurements and the system introduce. Thus, although the system state is hidden the human can still get a good estimate for it which is represented by a belief state.

The Kalman Filter includes two phases both of which result in a new belief.

Prediction The prediction phase, where an estimate of the state (ball position) is generated using input data, the system model and prior measurements.

Update The update phase, where the estimate of the state given by the prediction is updated using a measurement (observation of the ball).

The Kalman gain $\mathbf{K}$ is computed to determine how much weight will be given to the estimated state and how much to the measured one. The update phase will improve the estimate of the state but it is left out if there is no observation of the ball (i.e. no measurement), so if the ball is outside the human's field of view the human has to rely on predictions alone. This limitation only applies for the complex model since the simplified model includes an unlimited field of view and the human is able to observe the ball at every time step independent of its position.

Since in case of the non-linear complex model only the linear movement of the ball must be considered for state prediction (because the human's position is assumed to be observed perfectly) the standard Kalman Filter

$$
\begin{aligned}
\tilde{\boldsymbol{\mu}}_t &= \mathbf{A}\boldsymbol{\mu}_{t-1} + \mathbf{B}\mathbf{u}_{t-1} + \mathbf{g} && \\
\tilde{\boldsymbol{\Sigma}}_t &= \mathbf{A}\boldsymbol{\Sigma}_{t-1}\mathbf{A}^\mathsf{T} + \mathbf{R} && \text{prediction} \\
\mathbf{K}_t &= \tilde{\boldsymbol{\Sigma}}_t\mathbf{C}^\mathsf{T}(\mathbf{C}\tilde{\boldsymbol{\Sigma}}_t\mathbf{C}^\mathsf{T} + \mathbf{Q})^{-1} && \text{Kalman gain} \\
\boldsymbol{\mu}_t &= \tilde{\boldsymbol{\mu}}_t + \mathbf{K}_t(\mathbf{y}_t - \mathbf{C}\tilde{\boldsymbol{\mu}}_t) && \\
\boldsymbol{\Sigma}_t &= (\mathbf{I} - \mathbf{K}_t\mathbf{C})\tilde{\boldsymbol{\Sigma}}_t && \text{update}
\end{aligned}
$$

with the covariances of the system and observation noise $\mathbf{R}$ and $\mathbf{Q}$, suffices to gain the belief state[1] $(\hat{\boldsymbol{\Sigma}}_t, \hat{\boldsymbol{\mu}}_t)$ with the mean $\hat{\boldsymbol{\mu}}$ and the covariance $\hat{\boldsymbol{\Sigma}}$ for both models. The belief state is based on all observations $\mathbf{y}_{0:t}$ up to the current time step t, which is more precise than reducing the noise based on the current measurement alone. The mean vector $\hat{\boldsymbol{\mu}}_t$ of the belief state is used for the calculation of the motor commands instead of the system state $\mathbf{x}$, i.e.

$$\mathbf{u}_t = \mathbf{K}\boldsymbol{\mu}_t + \mathbf{k},$$

because using the mean is sufficient for a linear system and only the linear movement of the ball is considered for state estimation.

A distinctiveness of the observations in the complex model is that they are delayed by τ time steps to reflect the human latency, thus

$$\mathbf{y}_{t+\tau} = \mathbf{C}\mathbf{x}_t + \boldsymbol{\nu}_t.$$

[1] This belief state either reflects the prediction or the update since the latter is independent from the first. Thus it is marked as $(\hat{\boldsymbol{\Sigma}}_t, \hat{\boldsymbol{\mu}}_t)$ and not as $(\boldsymbol{\Sigma}_t, \boldsymbol{\mu}_t)$ to clarify that it may either represent the predicted *or* the updated belief.

Thus, for τ time steps from the initial one the human can only rely on his predictions acquired by the Kalman Filter and every measurement update will be taken into account τ time steps after the observation is made.

The initial mean $\hat{\boldsymbol{\mu}}_0$ of the belief state is the first estimate of the system state given by the initial system vector $\mathbf{x}_0$ before it is Gaussian distributed with $\mathbf{x}_0 \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$ to create different starting conditions for every trial. The initial covariance $\hat{\boldsymbol{\Sigma}}_0$ of the belief state defines the distribution over the initial state of the system and is chosen such that the following assumptions are met.

- The ball will be thrown from more or less the same position each time with the vertical offset differing less than the horizontal offset.
- The speed of the ball may vary significantly more than the position resulting in steeper, lower, farther or shorter ball flight trajectories. However, the initial angle is not assumed to change as much as the actual velocity of the ball, so the error of the initial velocity in x-direction has a larger value than the one of the vertical direction. The initial velocity for the y-direction in the complex model is not assumed to change much to ease the learning process.
- The human's initial position is assumed to remain almost the same for every trial as is the human's velocity or viewing direction (for the simple or complex model respectively) to ease the learning process.

The system and measurement covariances $\mathbf{R}$ and $\mathbf{Q}$ are defined such that the following assumptions are met.

- There is no system error to the human's position and viewing direction. Thus, it is assumed that the human always runs exactly where he wants to (except for the simplified model which incorporates a small error for the human velocity). There is also no system error to the ball's position since nothing can influence it directly. However, there is a small system error to the ball's velocity accounting for air resistance but not for strong influences on the ball's trajectory like strong wind or spin which are neglected to ease the learning process.
- The measurements of the ball's position and (in case of the simplified model) the human's position include an error to reflect the human's uncertainty.

For the exact values of the covariances $\boldsymbol{\Sigma}_0$, $\mathbf{R}$ and $\mathbf{Q}$ see Appendix A for the simplified model and Appendix B for the complex model. The values of these matrices were chosen to reflect reality with reasonable errors.

The Kalman Filter gives an estimation of the hidden system state $\mathbf{x}$ whenever the ball is in sight of the human and a prediction whenever it is not to compensate for the noise of both the system and the measurements. The belief state thus obtained is a statistically optimal estimation of $\mathbf{x}$.

3.2 Optimal Control Policy

To obtain optimal behaviour an optimal control policy is required by which the human sets his actions and updates his state to receive the optimal trajectory. Stochastic Optimal Control is investigated to gain the optimal control policy for the simplified model while for the complex model reinforcement learning is approached through CMA-ES.

3.2.1 Stochastic Optimal Control

In every time step, the human receives a reward given by the reward function

$$r(\mathbf{x}, \mathbf{u}) = -\mathbf{x}^T \mathbf{G} \mathbf{x} - \mathbf{u}^T \mathbf{H} \mathbf{u},$$

with the weights $\mathbf{G}$ and $\mathbf{H}$ as defined in Appendix A.

This reward function has to be maximized in order to achieve optimal behaviour and obtain the optimal control policy.

Since the system is linear, a Linear Quadratic Gaussian controller (LQG) is utilized to gain optimal behaviour based on Stochastic Optimal Control (Movellan, 2009) because it is efficient for finding the optimal solution for linear systems. To do so, a policy needs to be found that maximizes the expected reward over T time steps where T is the last time step when either the ball touches the ground or the human catches it. The expected reward for each state t is given by the value function $V(\mathbf{x}_t)$. The optimal value function $V^*(\mathbf{x}_t)$ maximizes the value function and therefore the expected reward. Thus, to obtain the optimal policy the optimal value function needs to be found.

Stochastic Optimal Control is derived from Bellman's principle of optimality (Movellan, 2009), which states that if there is an optimal solution for the time steps $t : T$ there must be an optimal solution for the time steps $t + 1 : T$. One can also say that optimal steps always consist of nothing but optimal sub-steps in regard to the reward function.

Bellman's principle of optimality specifies a value function

$$V(\mathbf{x}_t) = \underset{\mathbf{u}}{\operatorname{argmax}} \int p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t) V(\mathbf{x}_{t+1}) \, \mathrm{d}\mathbf{x}_{t+1} + r(\mathbf{x}_t, \mathbf{u}_t).$$

Solving the maximum (the derivation can be found in Appendix F) yields the optimal value function

$$V^*(\mathbf{x}_t) = -\mathbf{x}_t^T \mathbf{V}_t \mathbf{x}_t + 2\mathbf{x}_t^T \mathbf{v}_t + \text{terms independent of } \mathbf{x},$$

with

$$\begin{aligned} \mathbf{V}_t &= \mathbf{A}^T \mathbf{V}_{t+1} \mathbf{A} + \mathbf{R} - \mathbf{K} \mathbf{V}_{t+1}^T \mathbf{A}, \\ \mathbf{v}_t &= \mathbf{r}_t - \mathbf{A}^T (\mathbf{V}_{t+1}^T \mathbf{g} - \mathbf{r}_t) - \mathbf{K} (\mathbf{V}_{t+1}^T \mathbf{g} - \mathbf{r}_t) \end{aligned}$$

and with

$$\mathbf{K} = \mathbf{A}^T \mathbf{V}_{t+1} \mathbf{B} (\mathbf{B}^T \mathbf{V}_{t+1}^T \mathbf{B} + \mathbf{Q})^{-1} \mathbf{B}^T,$$

assuming $\mathbf{V}_{t+1}$ is quadratic (which is guaranteed for an LQG) otherwise the derivation would be wrong. The optimal policy along with the optimal gains is given by

$$\begin{aligned} \mathbf{u}_t^* &= -(\mathbf{B}^T \mathbf{V}_{t+1}^T \mathbf{B} + \mathbf{Q})^{-1} \mathbf{B}^T (\mathbf{V}_{t+1}^T (\mathbf{A}\mathbf{x}_t + \mathbf{g}) + \mathbf{r}_t) \\ &= \mathbf{K}_t \mathbf{x}_t + \mathbf{k}_t, \end{aligned}$$

which is the PD-Controller introduced for the system model with the gains

$$\begin{aligned} \mathbf{K}_t &= -(\mathbf{B}^T \mathbf{V}_{t+1}^T \mathbf{B} + \mathbf{Q})^{-1} \mathbf{B}^T \mathbf{V}_{t+1}^T \mathbf{A}, \\ \mathbf{k}_t &= -(\mathbf{B}^T \mathbf{V}_{t+1}^T \mathbf{B} + \mathbf{Q})^{-1} \mathbf{B}^T \mathbf{V}_{t+1}^T \mathbf{g} - (\mathbf{B}^T \mathbf{V}_{t+1}^T \mathbf{B} + \mathbf{Q})^{-1} \mathbf{B}^T \mathbf{r}_t. \end{aligned}$$

The ball flight needs to be simulated once to compute the gains for this policy because the values of $\mathbf{K}_t$ and $\mathbf{k}_t$ depend on the expected reward for the next time step $t+1$, which can only be obtained by running a simulation with the ball and the same parameters that are used for the later run involving the human.

Stochastic Optimal Control provides an optimal policy with regard to the assumed reward function which lets the human catch the ball reliably as can be seen in Figure 4.4.

3.2.2 CMA-ES

A Linear Quadratic Gaussian controller as utilized for the simplified model requires a linear system and thus cannot be used for the complex model except when applying linearisation, but then only locally optimal solutions could be found. Also, since the noise involved in the model creates new scenarios for every trial reinforcement learning is required to achieve meaningful results for the optimal control policy (i.e. a general policy leading to a catch).

Therefore, CMA-ES (Hansen and Ostermeier, 1996, Heidrich-Meisner and Igeltest, 2009) is utilized to learn the optimal policy because it works for non-linear systems, is comparably simple and does not include approximations that could lower the quality of the control policy. Since CMA-ES is an evolution strategy it maintains a distribution of candidate solutions for the control problem which evolves over time in regard to the more successful solutions. In reference to the ball catching scenario a candidate solution consists of multiple trials in which the human tries to catch the ball, all randomized with a different seed. The successfulness of a solution is derived from the fitness function f, which in this case averages over the accumulated reward

$$r(\mathbf{x}) = \begin{cases} 0, & \text{if } z_B > 0 \\ -((x_B - x_H)^2 + (y_B - y_H)^2), & \text{else.} \end{cases}$$

of all trials to receive the most successful solution with the smallest variance and the highest reward.

The initial exploration rate σ for finding new candidate solutions is given by $\sigma = 0.01$. The search space dimension is given by the number of iterations. In each iteration the most successful solutions are chosen. To maximize the likelihood of previously successful solutions the mean and covariance matrix of the distribution are updated in each iteration. The

current mean of the distribution is the current solution for the optimal control problem. To obtain the first mean a minimum and maximum range is given for each parameter to optimize.

For simplification the trials are randomized with the same seed for each iteration starting at 0 and incrementing by one for each trial. The effect of this simplification should be overfitting with good training results for few trials with many iterations because the optimal policy will be easily fitted to the repeating scenarios and specialize on them resulting in bad results for the test scenarios with fully random seeds and as such mostly very different noise characteristics from the training data.

To ease the challenge of finding an optimal control policy the number of parameters (features) for optimization is reduced by transforming the $\mathbf{x}$ vector such that the human is the center of the new coordinate system and the new x-axis is the viewing direction of the human. Thus, the new feature vector $\mathbf{x}'$ is given by

$$\mathbf{x}' = \begin{bmatrix} d_{IH} & \delta_{IH} & d_{BH} & \delta_{BH} & \alpha & \beta & \psi \end{bmatrix}^\mathsf{T}$$

with

- the distance d_{IH} between the predicted impact position of the ball and the position of the human,
- the angle δ_{IH} between the human's viewing direction and the predicted impact position of the ball,
- the distance d_{BH} between the ball position and the human position,
- the angle δ_{BH} between the human's viewing direction and the ball position,
- the elevation angle α,
- the horizontal angle β between the initial ball position, the human and the ball and
- the bearing angle ψ.

The number of features which must be optimized is given by $\mathbf{K}$ ($2 \times$ the number of features) and $\mathbf{k}$ (2×1) for the optimal policy

$$\mathbf{u} = \mathbf{K}\mathbf{x} + \mathbf{k}.$$

Furthermore, to simplify the learning process for different behaviour like running backwards while watching the ball and running forwards at maximum speed the optimal policy is used as an upper level gating policy for two option policies

$$\mathbf{u}_1 = \mathbf{K}_1\mathbf{x} + \mathbf{k}_1,$$
$$\mathbf{u}_2 = \mathbf{K}_2\mathbf{x} + \mathbf{k}_2,$$

with the gains $\mathbf{K}_1$, $\mathbf{k}_1$, $\mathbf{K}_2$ and $\mathbf{k}_2$. The gating policy switches between the two option policies which are basically simple policies but using two of them makes it easier to learn different behaviour because each of them can optimize on very different and probably even contradicting properties like running forwards at top speed or running slowly backwards watching the ball. The best time point for the switching must be learned. The gating policy includes the features for the two option policies and a switching time, which determines when the first policy is replaced by the second one and which is optimized along with the features. To receive the switching time relative to the predicted total number of time steps, its value is calculated by

$$t_{\text{switch}} = \text{sig}(t_{\text{switch}})T,$$

where sig is the sigmoid function used to ease the learning process for CMA-ES (by providing a smaller possible value range) and T is the predicted last time step calculated using only the predictions of the Kalman Filter for the input values of the ball.

The parameter ranges for the initial mean of the distribution for the simple and the upper level policy can be looked up in Appendix C. The learned mean as a result of the optimization has the same form as the initial mean and the gains $\mathbf{K}$ and $\mathbf{k}$ are retrieved from it as follows

$$\mu = [e_1, e_2, e_3 \ldots],$$
$$\mathbf{K} = \begin{bmatrix} e_1 & \cdots & e_{\#\text{features}} \\ e_{\#\text{features}+1} & \cdots & e_{2\#\text{features}} \end{bmatrix},$$
$$\mathbf{k} = \begin{bmatrix} e_{2\#\text{features}+1} \\ e_{2\#\text{features}+2} \end{bmatrix}$$

for the simple policy. For the upper level policy the above procedure will provide $\mathbf{K}_1$ and $\mathbf{k}_1$, repeating it for the next required number of elements will provide $\mathbf{K}_2$ and $\mathbf{k}_2$ and the switching time point is retrieved from the last element of the mean vector transformed with the above formula for t_{switch}.

CMA-ES provides an optimal policy given well defined features and an initial mean by which the human should be able to catch a ball in a testing scenario different from those he is trained on (i.e. with a different random seed). The results of this optimization are discussed in the next section.

4 Evaluation and Results

The aim of this thesis is to show that both reactive heuristics as well as a form of trajectory prediction can be reproduced for successful tests with the optimal policy when changing only noise and delay (which simulates the different ball flying characteristics). Thus, it should be seen in the results of the simulation that for sports like baseball where the noise plays a large role but the delay does not a relative heuristic is applied whereas for table tennis where the opposite is the case prediction is necessary and for tennis which is influenced by both noise and latency the human utilizes both behaviours.

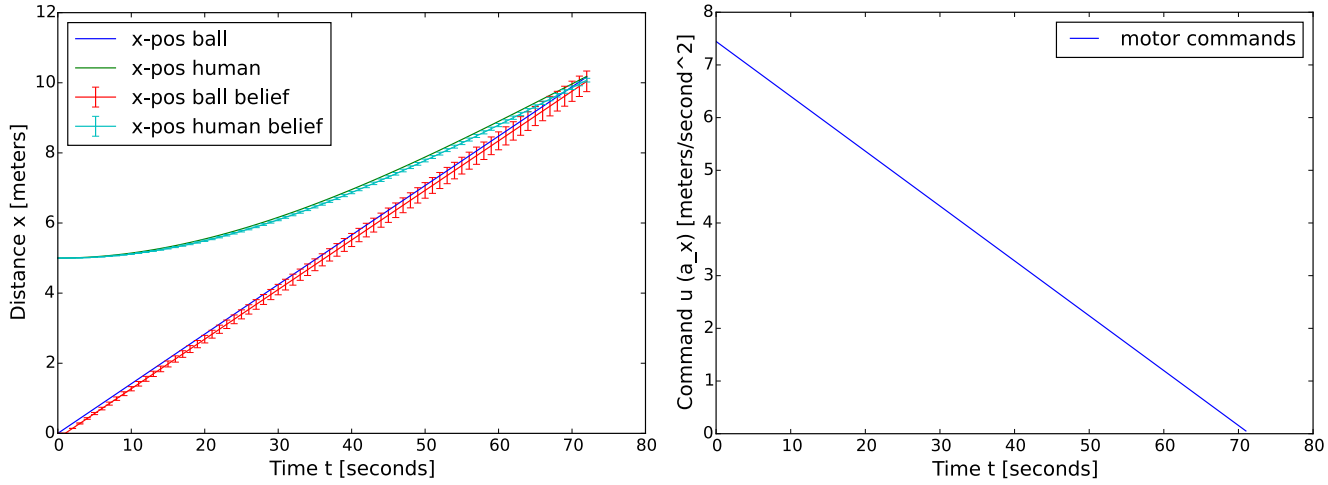
4.1 Simplified Model

The simplified model is only tested with regard to the TP and OAC heuristics since LOT and CBA require a three-dimensional state space (or at least a planar one for CBA). Also, the results gained from the simplified model might not be of high relevance due to the missing representation for the human latency, field of view and speed limitations.

4.1.1 Testing Trajectory Prediction

The assumptions made about Trajectory Prediction are that, while it is the optimal strategy for a catching scenario without or with a small amount of noise resulting in time-optimal behaviour, it will fail as soon as more noise is included into the model. To simulate the usage of TP, which only relies on the predictions based on the initial values, the measurement updates of the Kalman Filter are not included in this experiment.

The first test includes a small amount of noise, i.e. the initial noise is set to zero and only the system noise is taken into account. As expected, the human catches the ball (see Figure 4.1) being at the impact position when the ball arrives. The

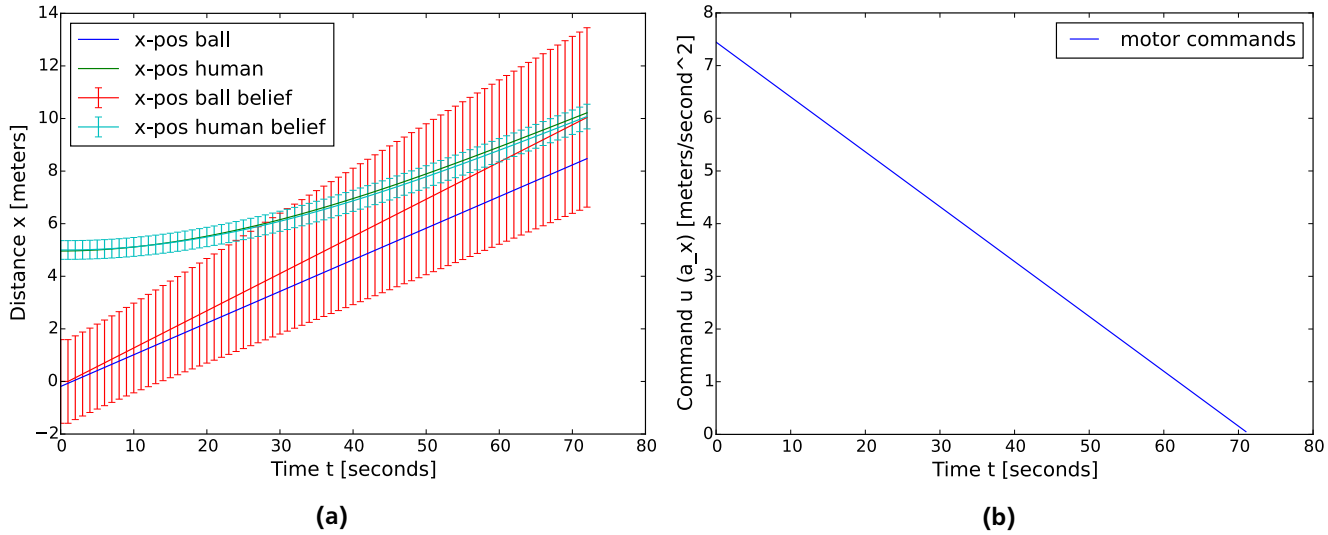


(a) (b)

Figure 4.1.: Example plots for very little (i.e. no initial noise) showing a successful catch despite no measurement updates are available. (a) This plot illustrates the catch where the human and the ball meet at the last time step. Because of the missing measurement updates and the system error the initially very low uncertainty (due to the missing initial noise) grows larger for every time step. (c) This plot shows the motor commands which indicate that the human is constantly decelerating.

interesting aspect to this result is that the human, though starting with a very high acceleration, constantly decelerates. This observation contradicts the assumption of time-optimal behaviour where the human would try to run to the impact position as fast as possible and not decelerate on the way.

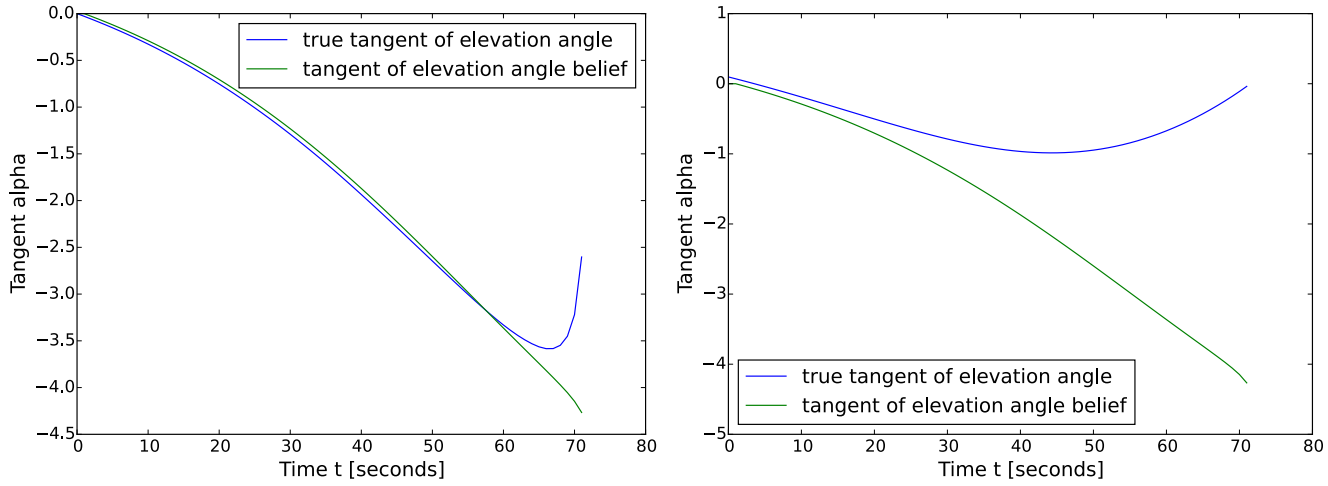
The second test includes the initial and system noise specified in Appendix A resulting in a large initial uncertainty which grows over time (see Figure 4.2). As expected, the human fails to catch the ball as soon as noise perturbs its trajectory and instead heads to the belief state he has of the ball's position (unless of course the ball happens to fly just as expected in which case the human would catch it by chance).



(a) **(b)**

Figure 4.2.: Example plots for a noisy catching scenario showing a failed catch because no measurement updates are available. (a) This plot illustrates the failed catch where the human reaches the belief state he has of the ball's position but the ball is somewhere else. (c) This plot shows the motor commands which indicate that the human is constantly decelerating.

Both tests for the successful and the failed catch show similar near-linear changes in the human's belief of the tangent of the elevation angle (see Figure 4.3). These results might indicate that the human tends to use the OAC theory (i.e. OAC describes his behaviour) instead of behaving time-optimal when the goal is not speed but precision.



(a) **(b)**

Figure 4.3.: Plots of the tangent of the elevation angle for the above scenarios. (a) The plot for the successful catch with zero initial noise. (b) The plot for the failed catch with initial noise.

4.1.2 Testing Optical Acceleration Cancellation

OAC states that the tangent of the elevation angle should change at a constant rate. To test this hypothesis, the tangent of the elevation angle is plotted over time for the tests of TP (see Figure 4.3) and for normal test runs including initial and system noise and measurement updates (see Figure 4.4). If OAC would apply, the tangent of the elevation angle would have to be linear over time.

While for the TP runs the change of the tangent is approximately linear the results for the normal runs (including measurement updates and noise) are not very promising. Though the human catches the ball he does not seem to try to maintain OAC by running so as to keep the tangent of the elevation angle changing linearly over time.

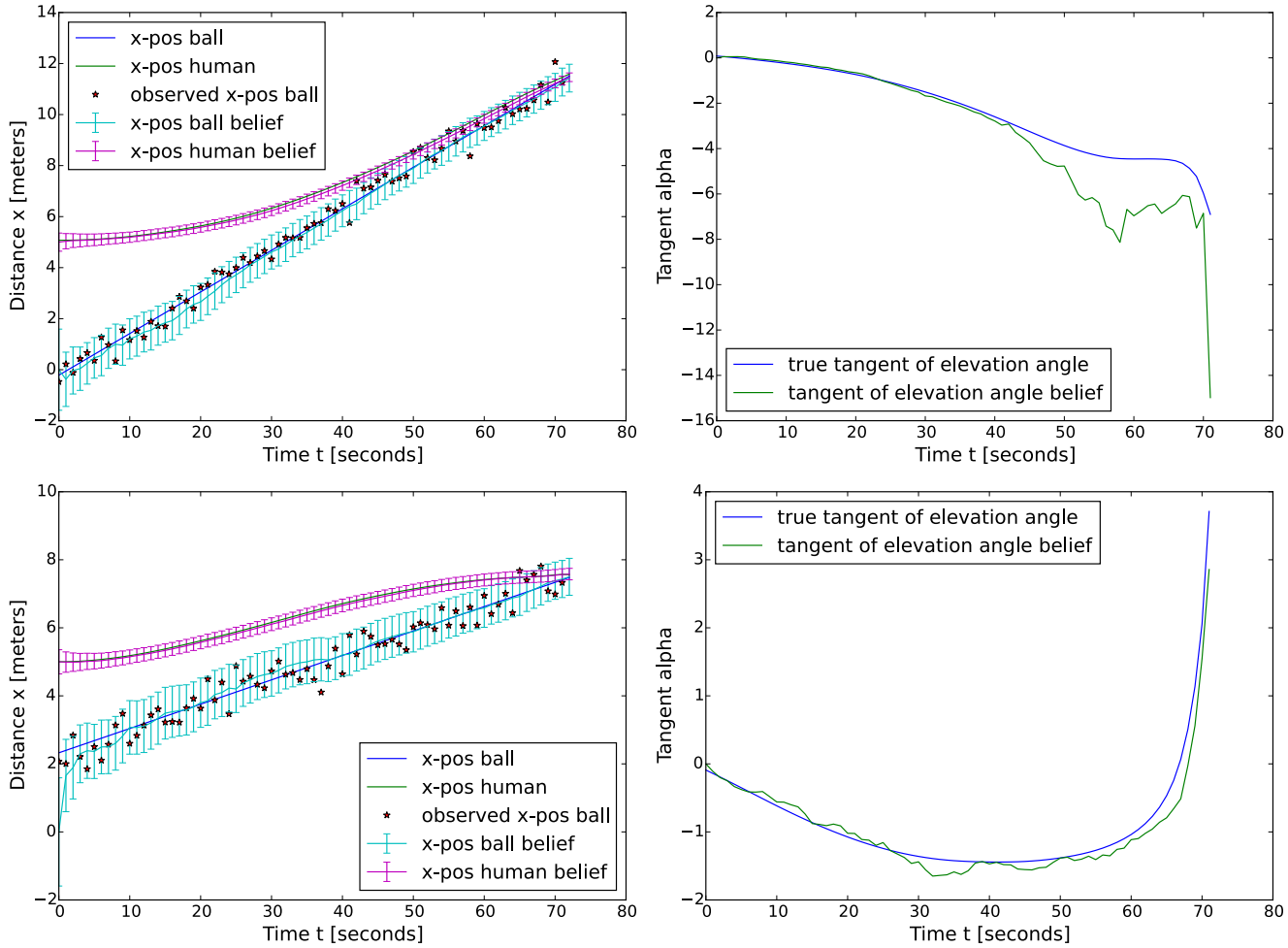


Figure 4.4.: Example plots showing successful catches for two different noisy test scenarios for which the measurement updates (i.e. the observations) are available. The policy is given by a LQG controller that is optimal for the given system.

4.2 Complex Model

The complex model is tested in regard to all heuristics to find an optimal policy which reproduces a heuristic but is also universal for different noise and delay settings.

Additional conditions for the tests are that the human always starts facing the ball which should ease the learning significantly because he immediately gets observations (only delayed by the latency). Also, the upper level policy implies that the human turns around in a single time step when the option policies are exchanged so the first option policy will be optimized for running backwards and watching the ball while the second one will be optimized for running forwards.

4.2.1 Testing Trajectory Prediction

The first heuristic to be considered is Trajectory Prediction. The assumptions made for TP are that, while it is the optimal strategy for a catching scenario without or with very low noise and a high latency resulting in a straight line trajectory to the impact position and probably time-optimal behaviour (the 2D results show that this is not necessarily the case), it will fail as soon as noise is included into the model. To simulate the usage of TP which only relies on the predictions based on the initial values the measurement updates of the Kalman Filter are not included in this experiment.

The first test includes a small amount of noise, i.e. the initial noise is set to zero and only the system noise is taken into account. As expected, the human catches the ball (see Figure 4.5) after little training (approximately 200 iterations of CMA-ES with one trial each) being at the impact position when the ball arrives.

As already observed for the simplified model, the human does not run at maximum speed all the time but speeds up backwards (see Figure 4.6) as opposed to the deceleration observed for the simplified model. This difference in behaviour is probably due to the motor command constraints introduced to the complex model. Moreover, the human

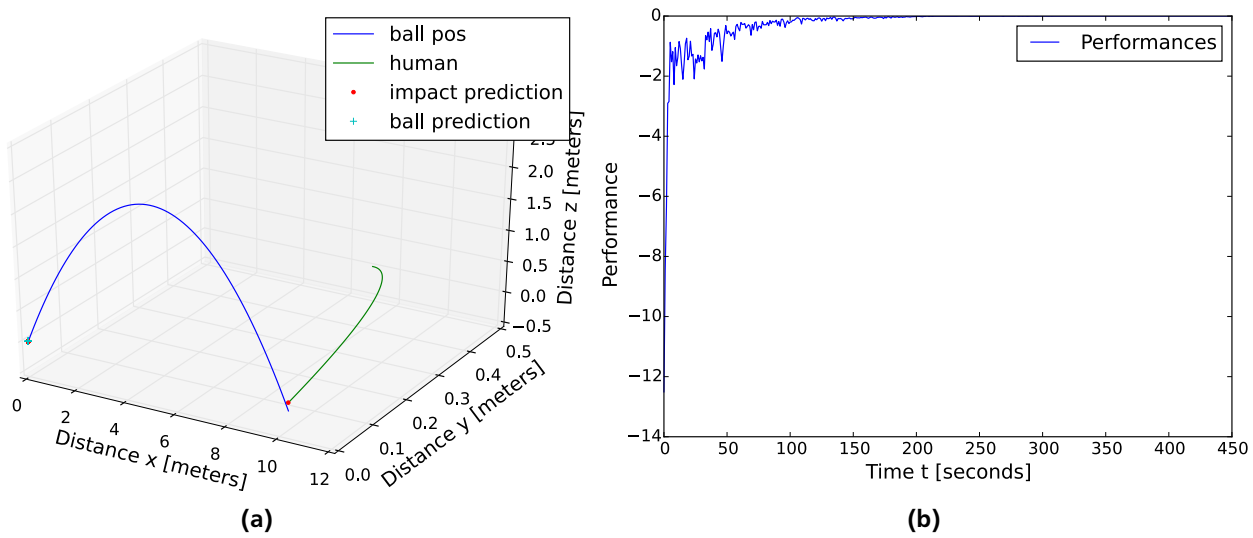


Figure 4.5.: Example plots for a small amount of noise showing a successful catch despite no measurement updates are available. (a) This plot shows the catch where the human and the ball meet after approximately ten meters of flight at the impact position marked by a red dot which is predicted by the human based on initial data. (b) This plot shows the performances for each iteration of CMA-ES given one trial. The training performances providing the data for this plot are each the mean performance of the current distribution of candidate solutions (i.e. the accumulated reward for all trials) found by CMA-ES. After approximately 200 iterations the human has learned how to run to always catch the ball.

does not move on a straight trajectory (see Figure 4.5) which again seems to contradict the assumption of time-optimal behaviour.

However, when taking a closer look at what the plots imply, the behaviour might be interpreted as time-optimal. The point is, the human is unable to change his direction instantly due to the rotational velocity constraints. Since he is not aligned with the ball he has to turn around to face the impact position. He could do so while standing still but he does not. Instead, he accelerates slowly while turning around, thus curving in the right direction (not away from the impact position due to a velocity that is chosen too high) while even getting a little closer to the impact position. After that the human behaves just as predicted running on a straight path at maximum speed.

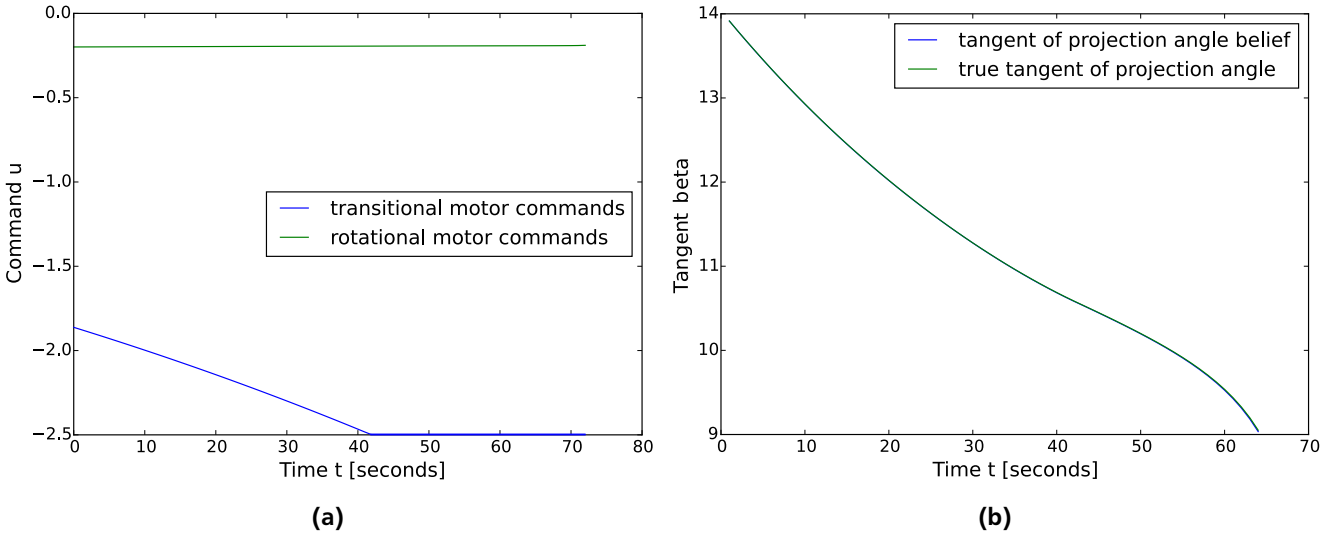
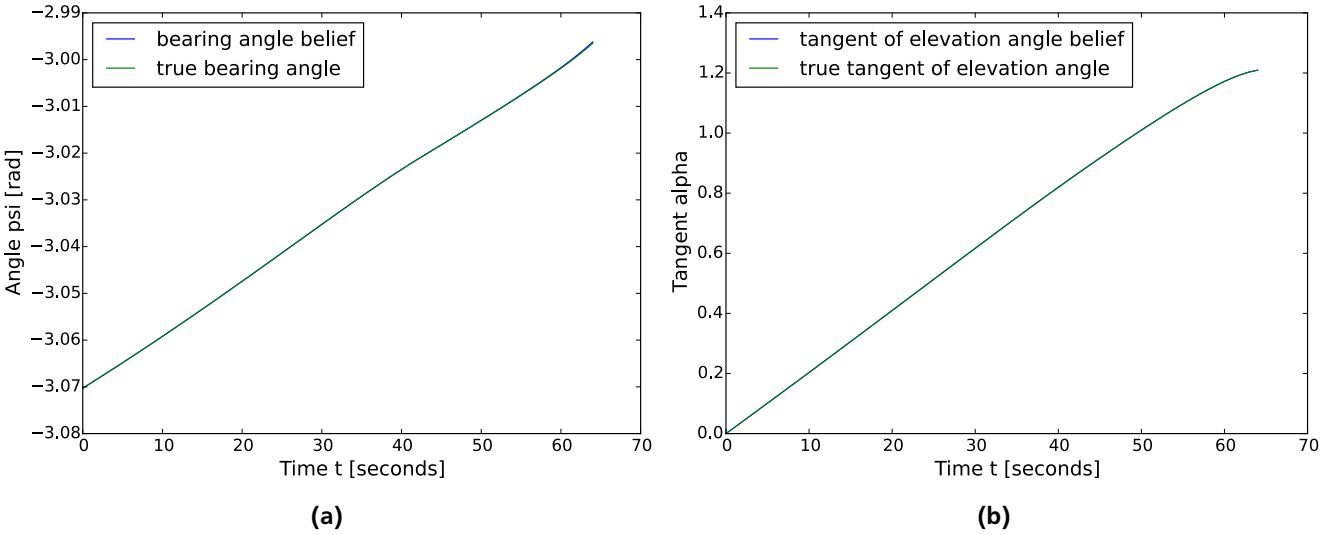


Figure 4.6.: Additional plots for the above scenario. (a) This plot illustrates the human's translational and rotational speed the first of which increases over time until it reaches the maximum. (b) This plot shows the tangent of the projection angle for LOT which does not stay constant.

This is all the more interesting because the plots regarding OAC and CBA strongly suggest that the human follows the OAC/CBA strategy (see Figure 4.7).



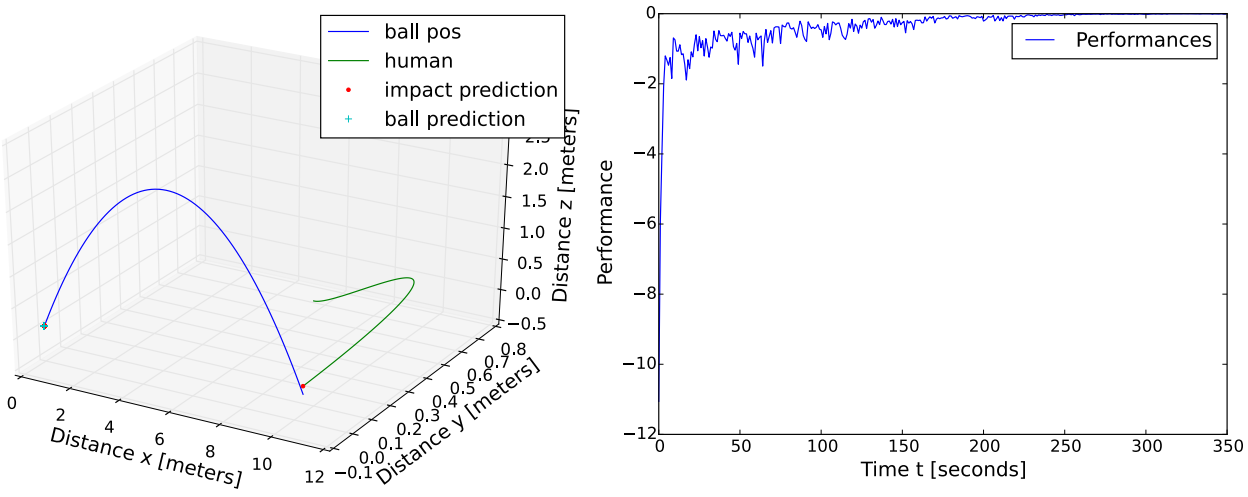
(a) (b)

Figure 4.7.: Additional plots for the above scenario. (a) This plot shows the bearing angle for CBA which remains approximately constant. (b) This plot shows the tangent of the elevation angle for OAC which changes constantly over time.

The bearing angle is kept nearly constant, changing only a little over time and the change of the tangent of the elevation angle is constant but for a very slight deviation at the end. This would imply that at least for scenarios with a small amount of noise the OAC/CBA strategy is time-optimal. It could also imply that TP is actually represented by OAC and CBA for a small amount of noise and available predictions for the ball's movements.

However, there is no evidence provided yet that the human uses LOT since the tangent of the projection angle is far from constant in Figure 4.6.

The same observations regarding the different heuristics apply for the second test (see Figure 4.8) which also includes a small amount of noise (i.e. no initial noise) but a high latency of 350 ms instead of the 180 ms used for the standard tests. The results are very similar to those with standard latency, but this time the human runs backwards at full speed from the beginning (see Figure E.1 for the plots showing the motor commands and the bearing angle for CBA and tangent for OAC and LOT).



(a) (b)

Figure 4.8.: Example plots for a small amount of noise but high latency showing a successful catch despite no measurement updates are available. (a) This plot shows the catch where the human and the ball meet after approximately ten meters of flight at the impact position (red dot) which is predicted by the human based on initial data. (c) This plot shows the performances for each iteration given one trial. After approximately 300 iterations the human always catches the ball.

For a higher latency it takes longer until the human learns the optimal policy for a correct catch (approximately 300 iterations of CMA-ES with one trial each) due to the added difficulty of compensating for the delay in the observations.

Though still an optimal policy can be found for which the human catches the ball high latency can indeed significantly complicate the task requiring about 100 additional iterations even when there is no initial noise and the impact position remains roughly the same for each trial.

The last test for TP includes the initial and system noise specified in Appendix B resulting in a large initial uncertainty which grows larger over time (see Figure 4.9). As expected, the human fails to catch the ball as soon as noise perturbs its trajectory. For a plot of the uncertainty of the human and the performance see Figure E.2. The only heuristic which might describe this scenario is OAC with a constant rate of change for the first and larger part of the run. Both the bearing angle and the tangent of the projection angle are not constant.

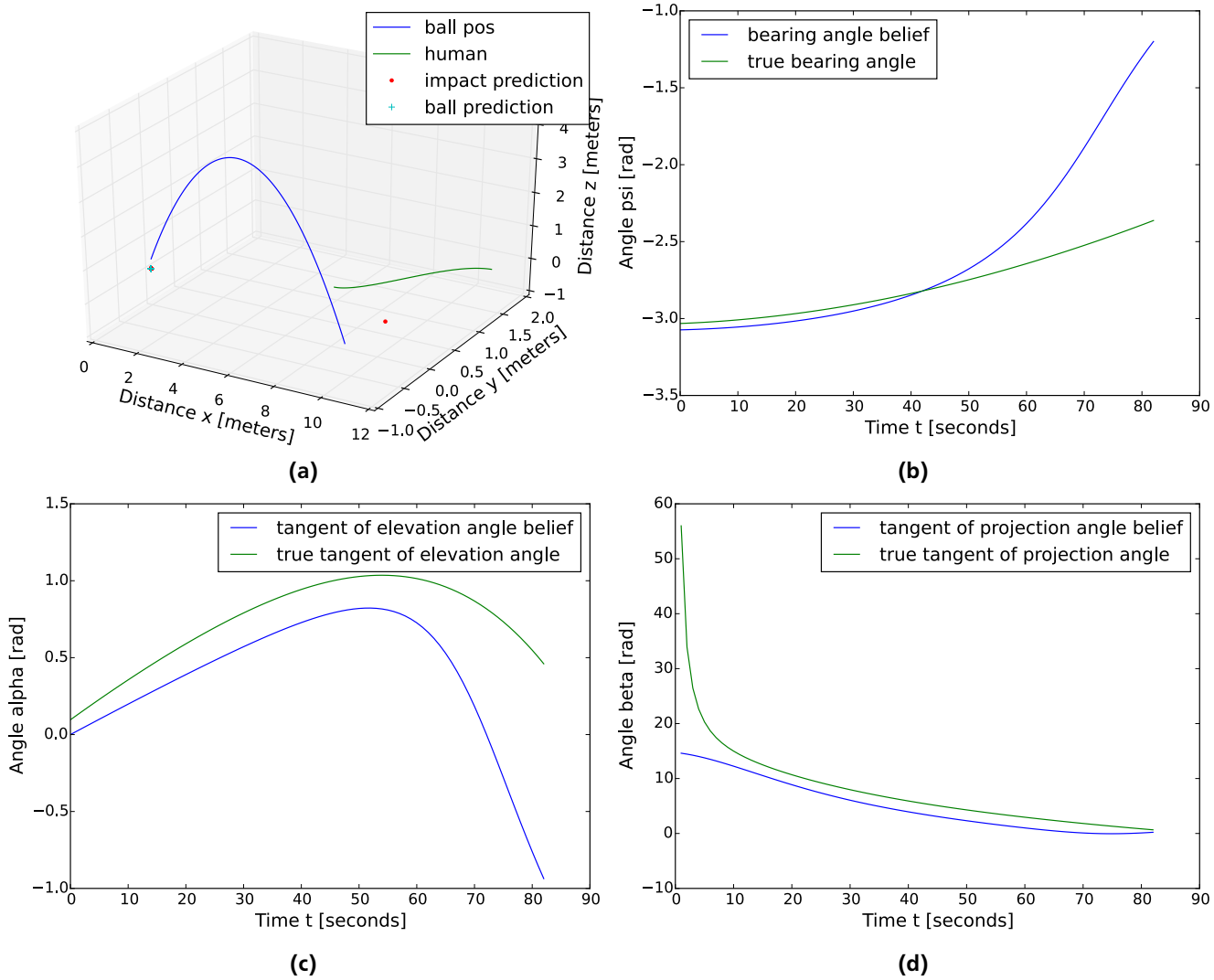


Figure 4.9.: Example plots for a noisy catching scenario showing a failed catch for which no measurement updates are available. (a) This plot shows the failed catch where the human misses the ball. (b) This plot shows the bearing angle of CBA which does not remain constant. (c) This plot shows the tangent of the elevation angle as described by OAC which increases constantly for most of the time. (d) This plot shows the tangent of the projection angle which is not constant.

4.2.2 Tests Including Measurements

In order to investigate the main assumption of this thesis that given the constrains of uncertainty and latency, the human behaviour (of following a heuristic) is optimal, the measurements are added to the model and the mean range as defined in Appendix C is used to learn the optimal settings for the simple and the upper level policy. It is assumed that the optimal policy will reproduce the behaviour of TP for high latency and of the reactive heuristics (OAC, CBA, LOT) for high noise and lower latency.

Different feature combinations are considered for optimization with CMA-ES. The features are chosen from the feature vector defined in 3.2.2 as

$$\mathbf{x}' = \begin{bmatrix} d_{IH} & \delta_{IH} & d_{BH} & \delta_{BH} & \alpha & \beta & \psi \end{bmatrix}^\mathsf{T},$$

with the distance d_{IH} between the predicted impact position of the ball and the position of the human, the angle δ_{IH} between the human's viewing direction and the predicted impact position of the ball, the distance d_{BH} between the ball position and the human position, the angle δ_{BH} between the human's viewing direction and the ball position, the elevation angle α, the horizontal angle β between the initial ball position, the human and the ball required for LOT and the bearing angle ψ.

From these features 14 different combinations are chosen which are regarded as the most interesting or promising ones and listed in Figure 4.10 since there is no possibility to test all feature combinations.

The combinations $(\delta_{BH}, \delta_{IH})$, (δ_{BH}, d_{BH}), (δ_{IH}, d_{IH}), (d_{BH}, d_{IH}) and $(\delta_{BH}, \delta_{IH}, d_{BH}, d_{IH})$ are chosen to get a good overview of the usefulness of the distance and angle to the impact position and to the ball relative to the viewing direction. Testing different combinations should help to find the most useful features.

The combinations (β, α), (ψ, α) and (β, ψ, α) are chosen to get a good overview of the usefulness of the different angles required by the heuristics. Since the elevation angle is required by both OAC and LOT and CBA is not regarded independent of OAC because it only considers lateral movement, α is present for all combinations. The first combination (β, α) is based on the required angles for OAC/CBA and the second (ψ, α) on the angles used by LOT (and OAC). The third combination (β, ψ, α) is chosen for completeness.

The remaining combinations are again combinations of those mentioned above. There is one to test the usefulness of providing all angles $(\delta_{BH}, \delta_{IH}, \beta, \psi, \alpha)$ which seems to make sense because all reactive heuristics rely on angles. However, the reward is given according to the distance from the impact position so different angle and distance combinations are included for all heuristics with one distance-angle combination and two heuristic (OAC/CBA, LOT/OAC) with two distance-angle combination $((\delta_{BH}, \beta, \psi, d_{BH}, \alpha)$, $(\delta_{IH}, \beta, \psi, d_{IH}, \alpha)$, $(\delta_{BH}, \delta_{IH}, \beta, d_{BH}, d_{IH}, \alpha)$ and $(\delta_{BH}, \delta_{IH}, \psi, d_{BH}, d_{IH}, \alpha))$. Finally, the whole combination of features is tested because it provides an unbiased overview (except for the bias included by choosing these features in the first place).

Each feature combination is tested for a number of trials ranging from 10 to 140 with a step width of 10 (smaller trial numbers would probably result in overfitting as shown in Figure E.3, larger ones would take very long) and iterations ranging from 200 to 600 with a step width of 100 (smaller iteration numbers would probably not give the human enough time to learn, larger ones would take very long). These numbers add up to 70 different learning scenarios for every feature combination and a total of 980 runs for all.

An overview over the training results obtained by these runs for both the simple and the upper level policy is shown in Figure 4.10. The training performances providing the data for these plots are each the mean performance of the best distribution of candidate solutions (i.e. the accumulated reward for all trials) found by CMA-ES in one run (consisting of e.g. 200 iterations and 20 trials).

The overview does not show a clear winner, especially since the training performances are not close to zero. Therefore, a closer look is taken at the best training result achieved during these runs for the simple and upper level policy respectively.

The best training result for the simple policy is given by $\boldsymbol{\mu}_{\text{best1p}}$ in Appendix D with a learning performance of approximately -5.736 after 600 iterations with 60 trials each for the feature combination $(\delta_{BH}, d_{BH}, \delta_{IH}, d_{IH})$. Compared to the average training performance of approximately -9.604 for 980 runs this is relatively good but as it reflects the negative quadratic distance to the ball when it touches the ground (i.e. the human is about 2.6 m from his goal compared to 3 meters average) it is not nearly good enough to count for a catch nor is it much better than the average. The learned control policy is tested for different noise and latency settings to find out how the human behaves for different catching scenarios.

The first tests includes standard initial and system noise (see Appendix B) and a latency of 180 ms. The results for these tests are very interesting. Though the human behaves not yet optimal he is never completely amiss and when he misses the ball it is often because he is not fast enough. This speed limitation might also account for the bad performance of the simple policy compared to the upper level policy in Figure 4.10 because the learned simple policy often delivers good results. Even more interestingly, the human actually seems to try to follow the OAC theory combined with CBA for lateral movement as indicated in Figure 4.11 and E.4 which shows another similar example and also in Figure 4.12 where the human tries to maintain a constant change of the tangent of the elevation angle and a constant bearing angle but this attempt fails because the ball is too fast for the human to catch.

Regarding OAC and CBA the policy yields quite good results so the same policy is tested for a scenario with zero initial noise and a higher latency of 230 ms instead of the usual 180 ms to represent the scenario of sports like table tennis. Since the human manages to be closer than 20 cm to the impact position (see Figure 4.13) this result is regarded as a catch indicating a first evidence for a universal catching theory, though with a small latency change only. The results for

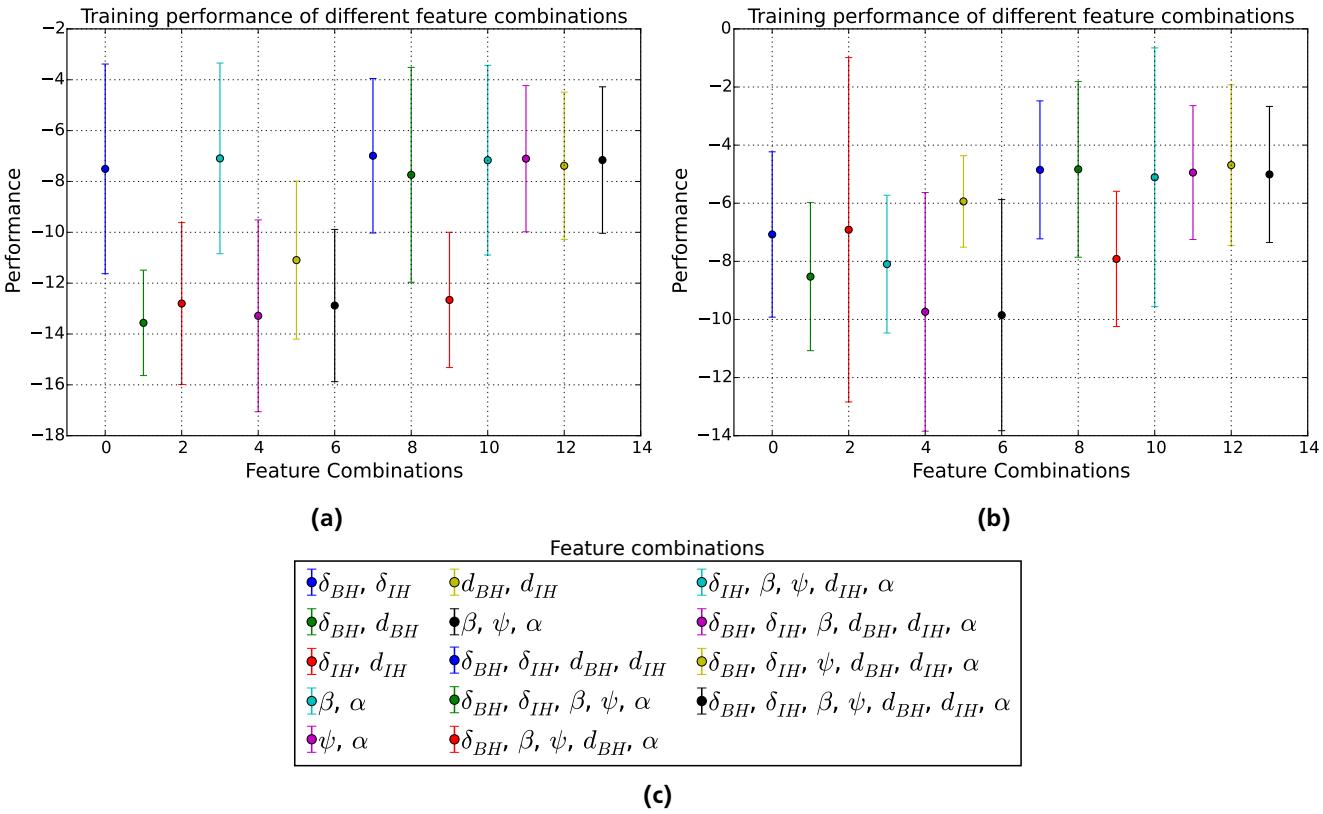


Figure 4.10.: An overview over the training results for 980 different learning scenarios and 14 different feature combinations. The dot marks the mean of the obtained performance values and the lines mark the standard deviation. Note that the reward can never actually become positive thus a very small deviation that includes zero would be optimal. (a) The training performances for the simple policy showing large deviations on the data with a rather bad performance. (b) The training performances for the upper level policy with two option policies showing better results though the test deviation is still large. (e) The feature combinations for the upper plots arranged from top to bottom and left to right to map the data in the plots from left to right.

OAC, CBA and LOT are similar to those gained for the TP tests, i.e. the human utilizes OAC and keeps the bearing angle approximately constant but LOT does not apply.

The best training result for the upper level policy is given by μ_{best2p} in Appendix D with a learning performance of approximately -2.244 after 600 iterations with 40 trials each for the feature combination $(\delta_{BH}, \delta_{IH}, \psi, d_{BH}, d_{IH}, \alpha)$. Compared to the average training performance of approximately -6.677 for 980 runs this is relatively good but as it reflects the negative quadratic distance to the ball when it touches the ground (i.e. the human is about 1.5 m from his goal compared to 2.5 meters average) it is not nearly good enough to count for a catch.

Unfortunately the human seems mostly unable to catch the ball at all when utilizing the upper level policy gained from the test runs but he does seem to try to follow the heuristics (see Figure 4.14). For an example of a successful catch see Figure E.7

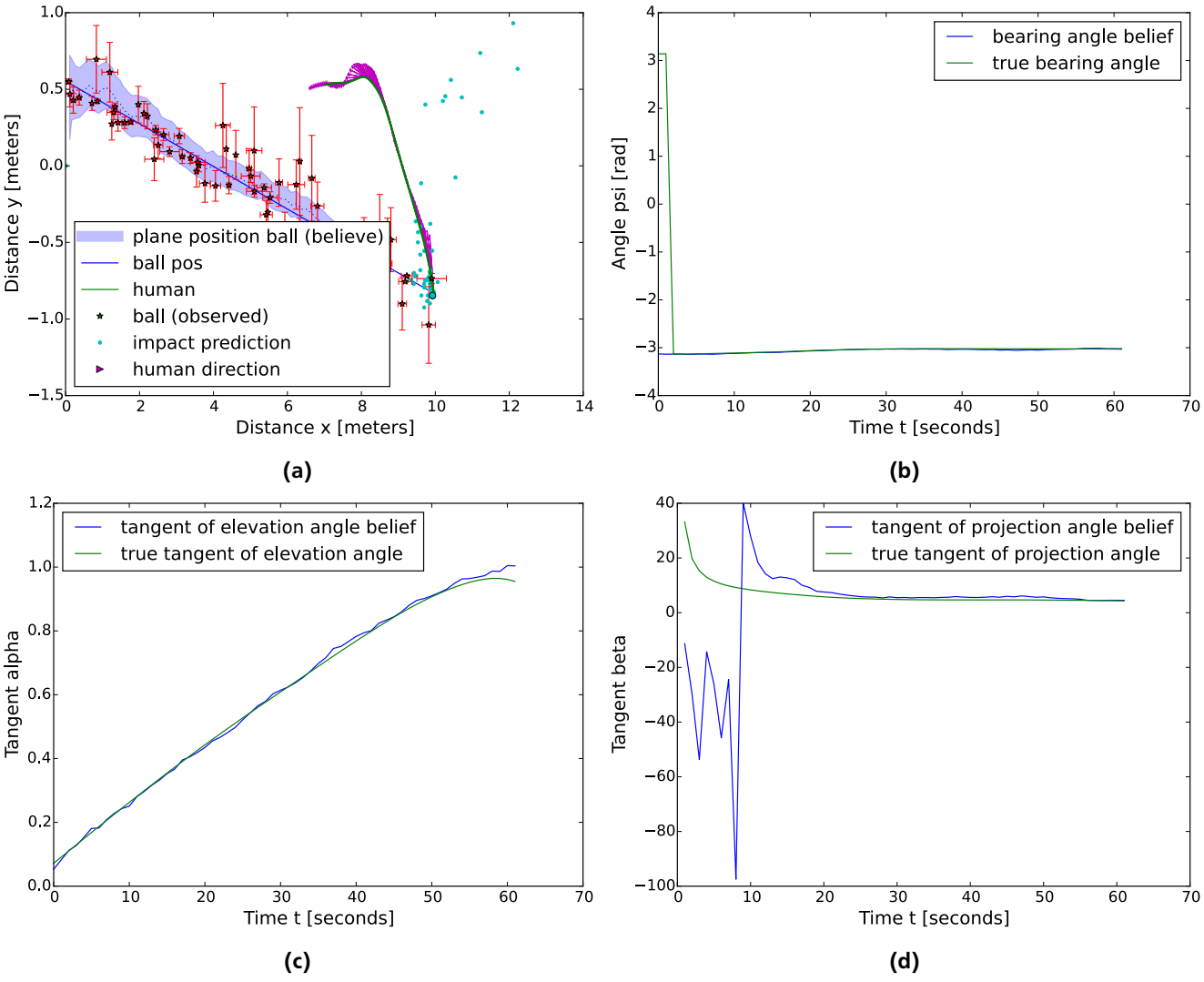


Figure 4.11.: Example plots for a catch with the simple policy. (a) This plot indicates the catch and shows that due to the backwards starting position (indicated by the backwards arrows) the human is always able to observe the ball. (b) The bearing angle stays approximately constant as predicted by CBA, except for the jump in the beginning which is a small change due to the periodicity of angles (this jump is optimized away for the learning process). (c) The tangent of the elevation angle increases approximately linearly as predicted by OAC. (d) The tangent of the projection angle is not constant as should be the case for LOT.

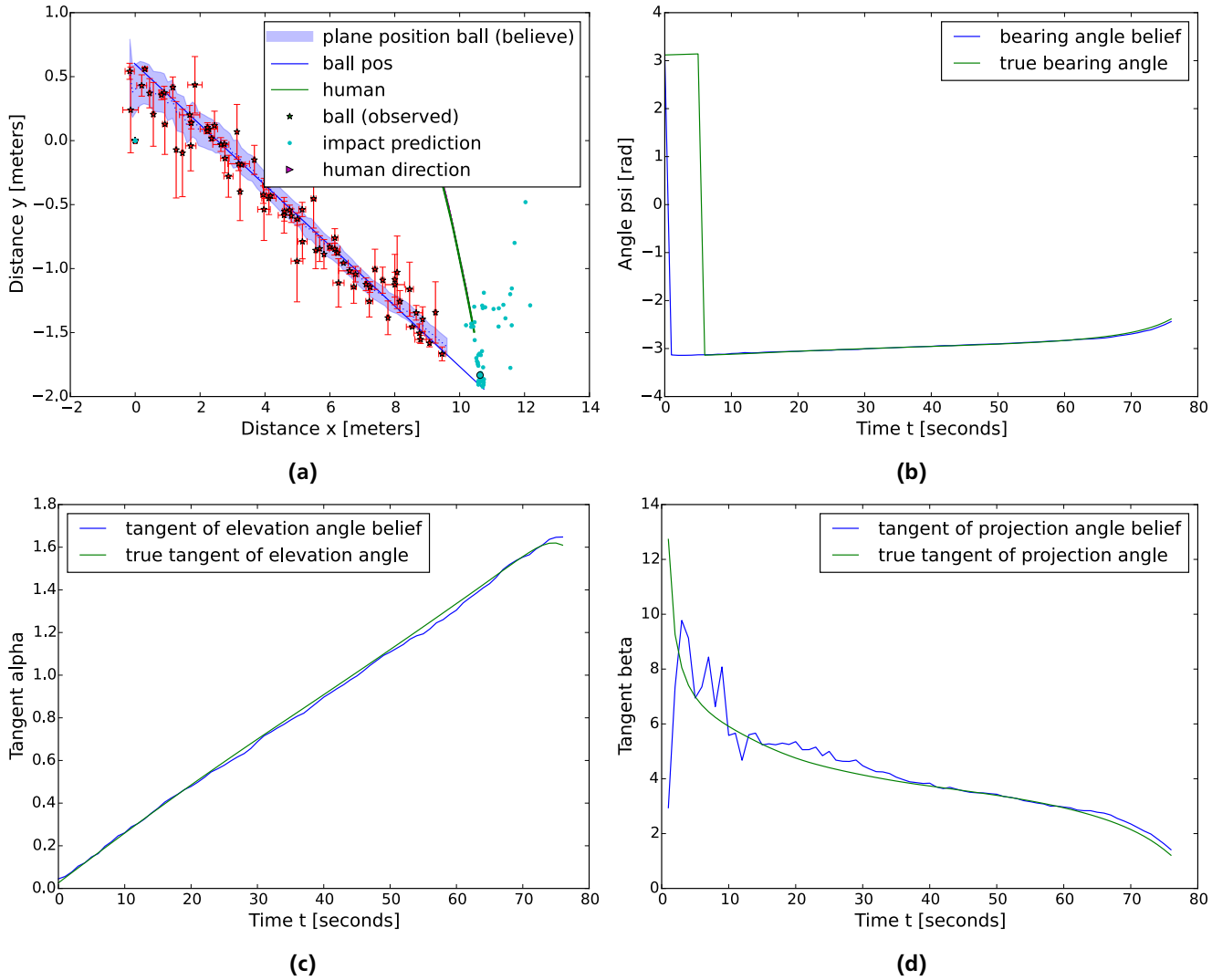


Figure 4.12.: Example plots for a failed catch where the ball is simply too fast for the human to catch though he gets very close. (a) This plot indicates the catch and shows that due to the backwards starting position (indicated by the backwards arrows) the human is able to observe the ball until it flies over his head. (b) The bearing angle stays approximately constant as predicted by CBA (c) The tangent of the elevation angle increases constantly as predicted by OAC. (d) The tangent of the projection angle is not held constant as should be the case for LOT.

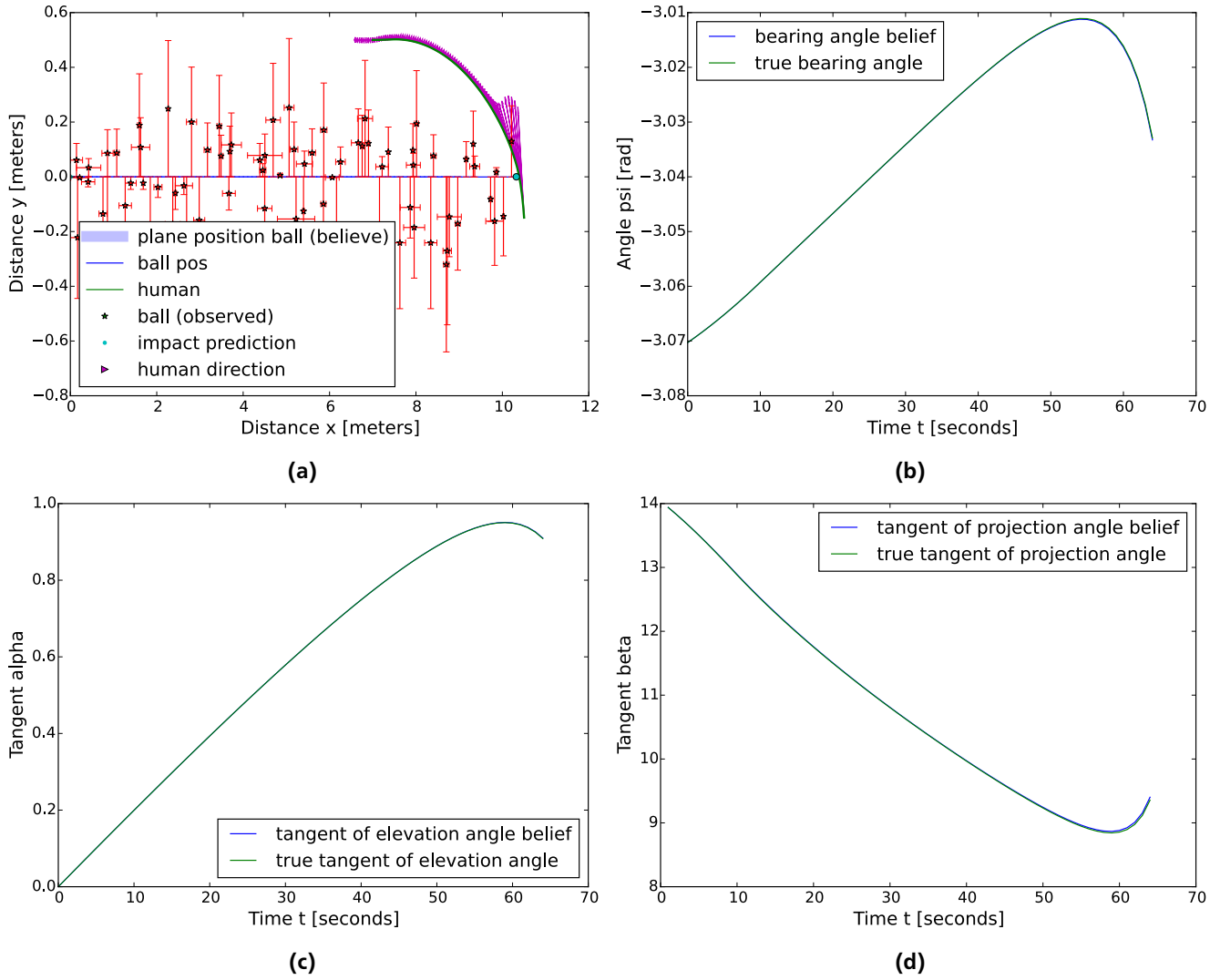


Figure 4.13.: Example plots for a successful catch where the policy adapts to the high latency test for no initial noise. (a) This plot indicates the catch. (b) The bearing angle stays approximately constant until the human runs ahead of the ball. (c) The tangent of the elevation angle changes linearly over time until the human runs ahead of the ball. (d) The tangent of the projection angle changes significantly as already observed for the TP tests.

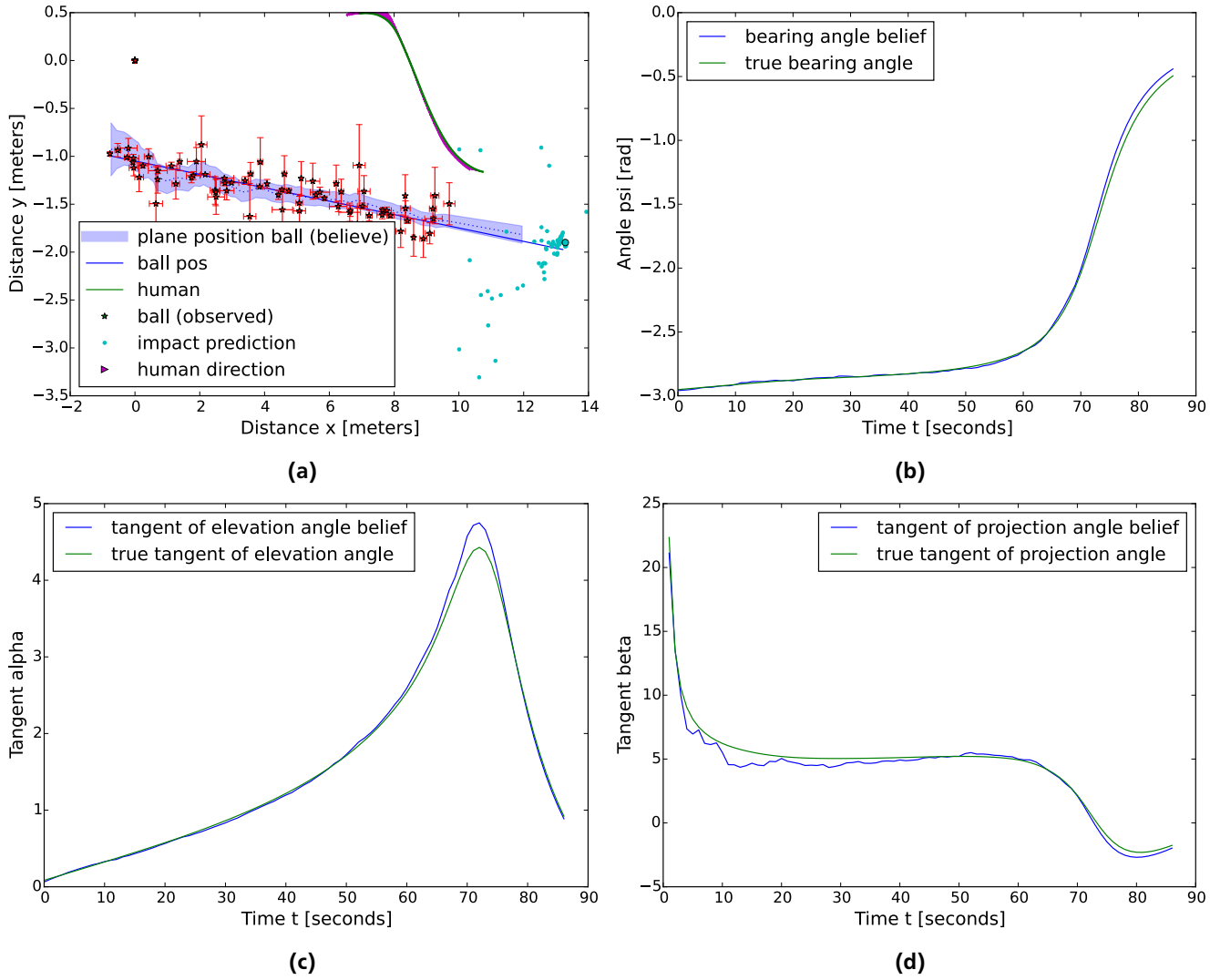


Figure 4.14.: Example plots for a failed catch with the upper level policy where the bearing angle and the tangent of the elevation angle show the same characteristics as described in Figure 4.12 for the simple policy. (a) This plot indicates the failed catch which is probably due to a large distance between the human initial and the impact position. (b) The bearing angle stays approximately constant until the ball flies ahead of the human. (c) The tangent of the elevation angle changes linearly over time until the ball flies ahead. (d) The tangent projection angle does not stay constant.

5 Discussion and Conclusion

The results from the previous chapter provide some interesting insights into human ball catching behaviour. The assumptions about TP failing for high noise and applying both for low noise and for low noise and high latency proves to be correct for both the simplified and the complex model. Additionally, for the complex model the human shows (approximately) time-optimal behaviour running at maximum speed on a straight path for most of the way except for the start when he has to turn around to face the impact position.

An unexpected discovery resulting from the TP tests is that the human still follows the OAC theory and approximately keeps the bearing angle of the CBA theory constant (though a slight increase can be observed for all tests without initial noise). This may imply that TP is actually a combination of OAC and (a variant of) CBA with the addition that predictions of the ball's position are available to the human. If this is the case and TP proves to be indeed time-optimal this would imply that OAC and CBA are optimal strategies for catching a ball. Further tests and research in this direction seem promising.

The observations also strengthen the belief of Gigerenzer et al. (Marewski, Gaissmaier, and Gigerenzer, 2009) that reactive policies are applied whenever possible. This observation may imply that the human tries to follow the heuristics because they help him catch the ball or they might be defined in such a way that they apply for most possible catches.

Nevertheless, the results for the simple policy where the ball is always in field of view of the human seem to be better than those for the upper level policy as long as the human is in range of the impact position. This observation suggests that the human relies very strongly on his vision of the ball and that it would probably be a better idea for future work to let the human run sideways (slower, but viewing the ball) and probably limit the field of view above the human.

Generally the learned simple policy produces good results which are mostly limited by scenarios with uncatchable balls. For future tests these scenarios should be prevented both for easier learning with CMA-ES and for gaining a better overview over the results. Doing so may be the fastest and most promising way based on the results gained in this thesis for finding the stable optimal policy for human ball catching.

The results strongly support OAC and mostly support CBA even for uncatchable fly balls where it seems that the human switches to another bearing angle (see Figure E.6) similar to the switching observed for dogs catching frisbees (Shaffer, Krauchunas, Eddy, and McBeath, 2004). LOT on the other hand does not seem to play a role for human ball catching since no scenario could be found where the tangent of the projection angle was held constant. These results agree with those found in a virtual reality experiment (Fink, Foo, and Warren, 2009).

Furthermore, evidence for a universal policy is indicated for the learned simple policy. Further tests in this direction should definitely be considered for future research.

Motor command punishments might also be considered because tests where the human starts near the impact position often show circling movements instead of slower movements which might be fixed with appropriate punishments.

Altogether the results are quite promising regarding the use of OAC and CBA and giving some evidence to Gigerenzer et al. as well as the hypotheses posed in this thesis. Further research could include more tests using policies with good results as starting ranges for CMA-ES. The latency and noise could be changed over time to reflect a ball flight more correctly and give the human the chance to look away for some time and still catch the ball looking back again when it is already quite near to him, which should significantly reduce the uncertainty.

For later tests with a stable policy one could also consider using high system noise to simulate a frisbee or wind blowing the ball away to test the policy for extreme conditions. Additionally, it would be interesting to fully randomize the training data when trying to find a universal optimal policy for different noise and latency settings.

Within the scope of this thesis, the optimal policy was found for the two-dimensional model and good results for the simple policy of the complex model were achieved which were mostly limited by the system constraints regarding the human's speed. Additionally, strong support for the reactive heuristics OAC and CBA was given and the assumptions regarding noise and latency for TP were reinforced. Finally, hints were found that OAC and CBA are optimal given the human constraints and that there is a common framework for human ball catching. Future work based on these results appears promising for the aim of finding a stable and universally adaptive optimal control policy.

Appendix

A Matrices for the Simplified Model

The system matrices **A**, **B** and the vector **g** are defined as

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \Delta t & 0 & 0 & 0 \\ 0 & 1 & 0 & \Delta t & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \Delta t^2 \\ \Delta t \end{bmatrix} \quad \mathbf{g} = \begin{bmatrix} 0 \\ -\Delta t^2 g \\ 0 \\ -\Delta t g \\ 0 \\ 0 \end{bmatrix}$$

with the time step $\Delta t = 0.02$.

The filter matrix **C** to extract the elements for observation from the state **x** is defined by

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

The reward weight matrix $\mathbf{G}_T$ for the state at the last time step is given by

$$\mathbf{G}_T = \begin{bmatrix} -c & 0 & 0 & 0 & c & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ c & 0 & 0 & 0 & -c & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The value c is an arbitrary reward. Choosing a large value like $c = 10^7$ gives good results in the case of the simplified model. The reward weight matrix **H** for the motor commands is given by

$$\mathbf{H} = \begin{bmatrix} -10^{-2} \end{bmatrix}$$

for every time step.

The initial variance $\hat{\mathbf{\Sigma}}_0$ of the belief state is a diagonal matrix with the diagonal entries

$$\hat{\mathbf{\Sigma}}_0 = \begin{bmatrix} 0.4, & 0.2, & 4, & 2, & 10^{-3}, & 10^{-3} \end{bmatrix}$$

determined through tests to reflect reality (see Section 3.1 for details).

The covariances **R** and **Q** are defined as diagonal matrices with the diagonal entries

$$\mathbf{R} = \begin{bmatrix} 0, & 0, & 10^{-5}, & 10^{-5}, & 0, & 10^{-8} \end{bmatrix}$$
$$\mathbf{Q} = \begin{bmatrix} 10^{-1}, & 10^{-1}, & 10^{-3} \end{bmatrix}$$

determined through tests to reflect reality (see Section 3.1 for details).

B Matrices for the Complex Model

The system matrices **A**, **B** and the vector **g** are defined as

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{B}_t = \begin{bmatrix} \Delta t \cos \phi_{H_t} & 0 \\ \Delta t \sin \phi_{H_t} & 0 \\ 0 & \Delta t \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \mathbf{g} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -g\frac{\Delta t^2}{2} \\ 0 \\ 0 \\ -g\Delta t \end{bmatrix}$$

with the time step $\Delta t = 0.02$.

The filter matrix **C** to extract the elements for observation from the state **x** is defined by

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

The initial variance $\hat{\mathbf{\Sigma}}_0$ of the belief state is a diagonal matrix with the diagonal entries

$$\hat{\mathbf{\Sigma}}_0 = \begin{bmatrix} 10^{-3}, & 10^{-3}, & 10^{-3}, & 0.4, & 0.4, & 0.2, & 4, & 0.8, & 2 \end{bmatrix}$$

determined through tests to reflect reality (see Section 3.1 for details).

The covariances **R** and **Q** are defined as diagonal matrices with the diagonal entries

$$\mathbf{R} = \begin{bmatrix} 0, & 0, & 0, & 0, & 0, & 0, & 10^{-5}, & 10^{-5}, & 10^{-5} \end{bmatrix}$$
$$\mathbf{Q} = \begin{bmatrix} 10^{-3}, & 10^{-3}, & 10^{-3} \end{bmatrix}$$

determined through tests to reflect reality (see Section 3.1 for details).

C Parameter Ranges for the CMA-ES Initial Mean

The parameter ranges for the simple policy are given by -2 for every entry in the minimum range vector and by 2 for every entry in the maximum range vector. The length of the range vectors is given by two times plus two the number of features. So in case only one component from the feature vector is used for CMA-ES the ranges are given by

$$\begin{aligned}\textbf{minrange} &= \begin{bmatrix}-2 & -2 & -2 & -2\end{bmatrix}\\ \textbf{maxrange} &= \begin{bmatrix}2 & 2 & 2 & 2\end{bmatrix}.\end{aligned}$$

The parameter ranges for the upper level policy including two option policies are given by -2 for every entry in the minimum range vector but for the last element which is the switching time given by -3 and by 2 for every entry in the maximum range vector but for the last element which is the switching time 3. The length of the range vectors is given by two times plus two the number of features plus one for the switching time. So in case only one component from the feature vector is used for CMA-ES the ranges are given by

$$\begin{aligned}\textbf{minrange} &= \begin{bmatrix}-2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -3\end{bmatrix}\\ \textbf{maxrange} &= \begin{bmatrix}2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 3\end{bmatrix}.\end{aligned}$$

D Best Mean Results

The best result for one policy with the standard range is given by

$$\mu_{\text{best1p}} = \begin{bmatrix} -18.57981202, & -19.57566965, & 11.73037107, & 4.92902204, & -0.09183829, \\ 8.13608147, & 0.16349904, & -0.03963051, & -4.33733074, & -25.08690407 \end{bmatrix}.$$

The best result for two policies with the standard range is given by

$$\mu_{\text{best2p}} = \begin{bmatrix} -1.84774988, & 1.01072449, & -2.26454588e{-}01, & -5.26885821e{-}02, \\ -3.97968646e{-}01, & 3.72432268e{-}01, & 1.40468831e{-}03, & 2.55451561, \\ -4.59887258e{-}01, & -1.01973255, & 9.45566616e{-}01, & -1.22885996, \\ 1.65524078e{-}01, & -2.06393210, & 1.85078633, & 2.75655908e{-}01, \\ -1.10855052e{-}01, & 1.25984555e{-}01, & -5.29628412e{-}01, & -1.72173970e{-}01, \\ 7.72064676e{-}02, & 2.13637975e{-}02, & 2.30850047e{-}01, & 2.24751491e{-}02, \\ 2.01708164e{-}02, & -2.51270448e{-}01, & 1.37881636e{-}01, & -8.17813369e{-}01, \\ 2.59739073e{-}01 & & & \end{bmatrix}.$$

E Example Plots

The following plots show the motor commands and OAC, CBA and LOT values for the experiment shown in Figure 4.8.

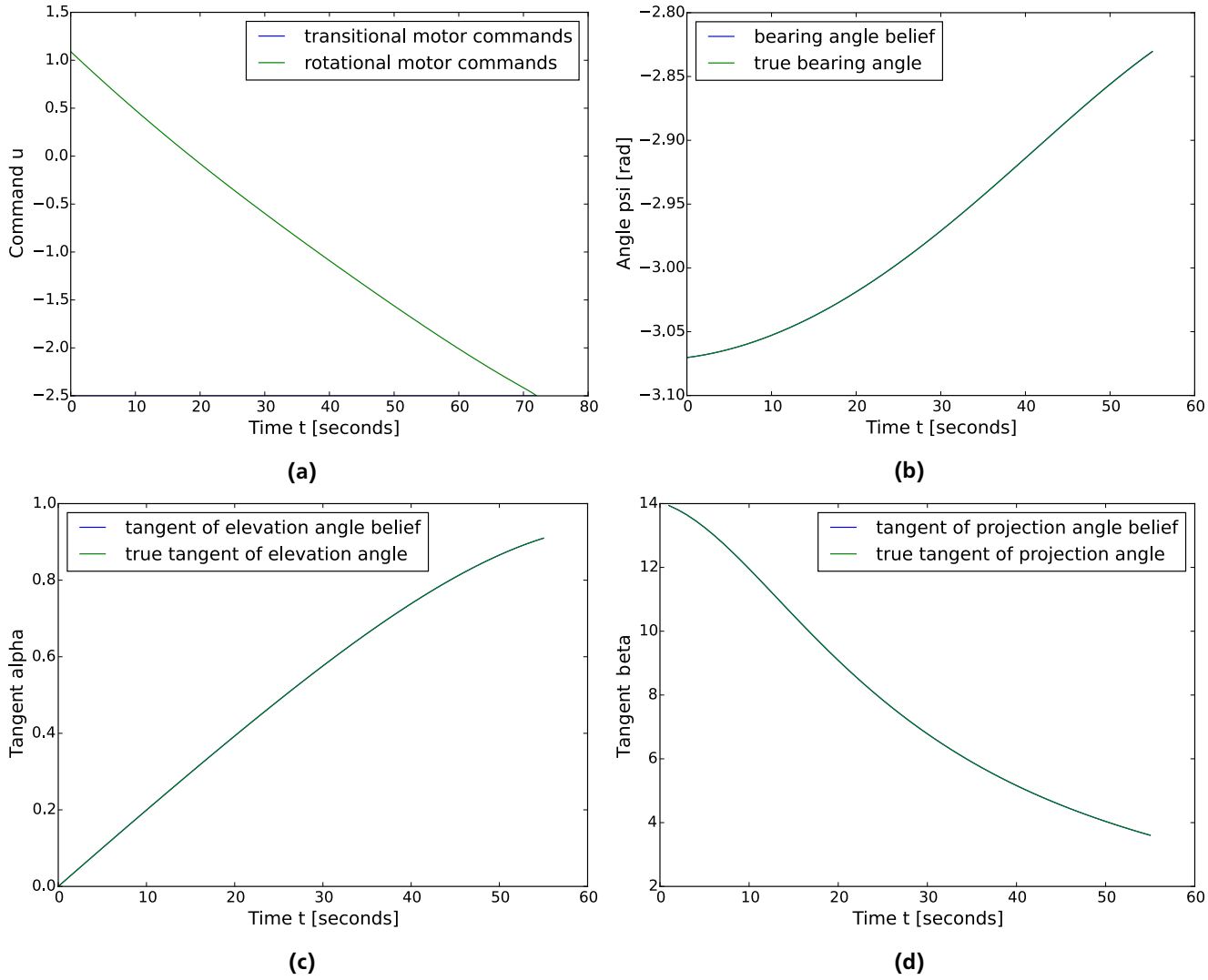
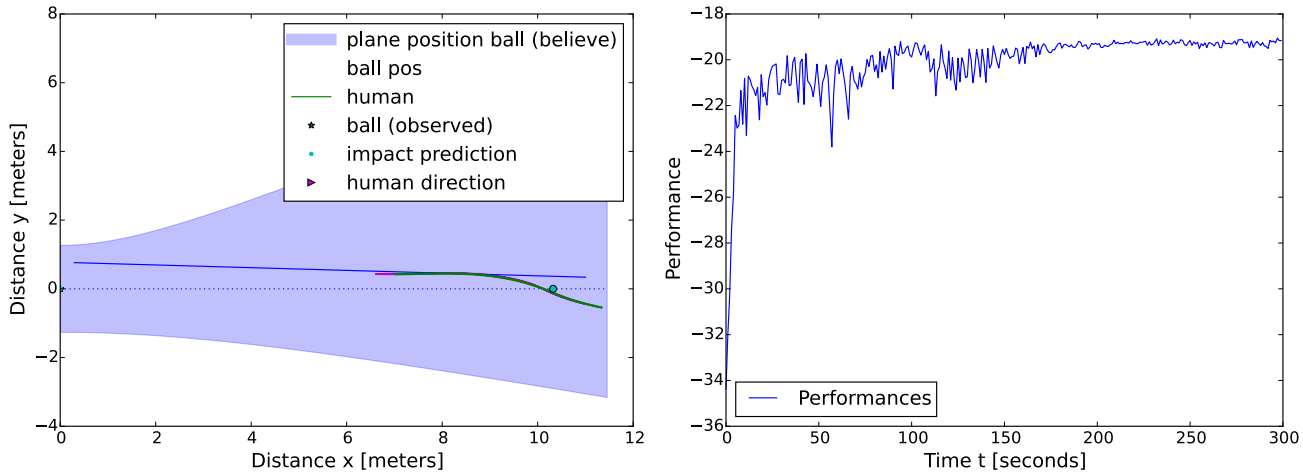


Figure E.1.: Example plots for a small amount of noise (no initial noise) but high latency and a successful catch despite no measurement updates are available. (a) This plot illustrates the human's translational and rotational speed the first of which is always at maximum backwards speed. (b) This plot shows the bearing angle of CBA which remains basically constant but increases a little. (c) This plot shows the tangent of the elevation angle as described by OAC which increases constantly over time. (d) This plot shows the tangent of the projection angle for LOT which decreases instead of staying constant.

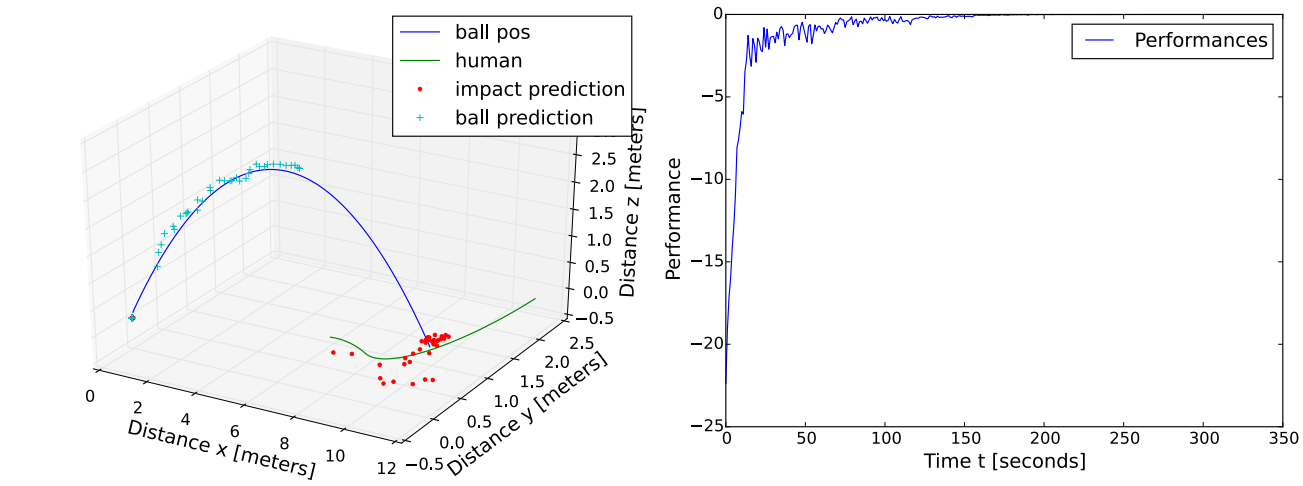
The following plots show the uncertainty and performance for the experiment shown in Figure 4.9.



(a) (b)

Figure E.2.: Example plots for a noisy catching scenario showing a failed catch for which no measurement updates are available. (b) This plot illustrates the human's uncertainty and blind movements. Because of the missing measurement updates the initially high variance to the belief grows even larger for every time step. (c) This plot shows the training performances for each iteration which is very bad and does not increase towards the end suggesting that a local maximum is reached.

The following is an example of overfitting for CMA-ES.



(a) (b)

Figure E.3.: Bad test results due to overfitting with one trial, 350 iterations and a training performance of -4.22649717232e-05. (a) The human misses the target relying on overfitted training data. (b) The training results show that the human learns very fast how to catch the ball when the random seed remains the same for each trial and only one trial is repeated very often.

The following is an example of a successful catch utilizing the best simple control policy found which shows similar results considering the usage of the heuristics as Figure 4.11.

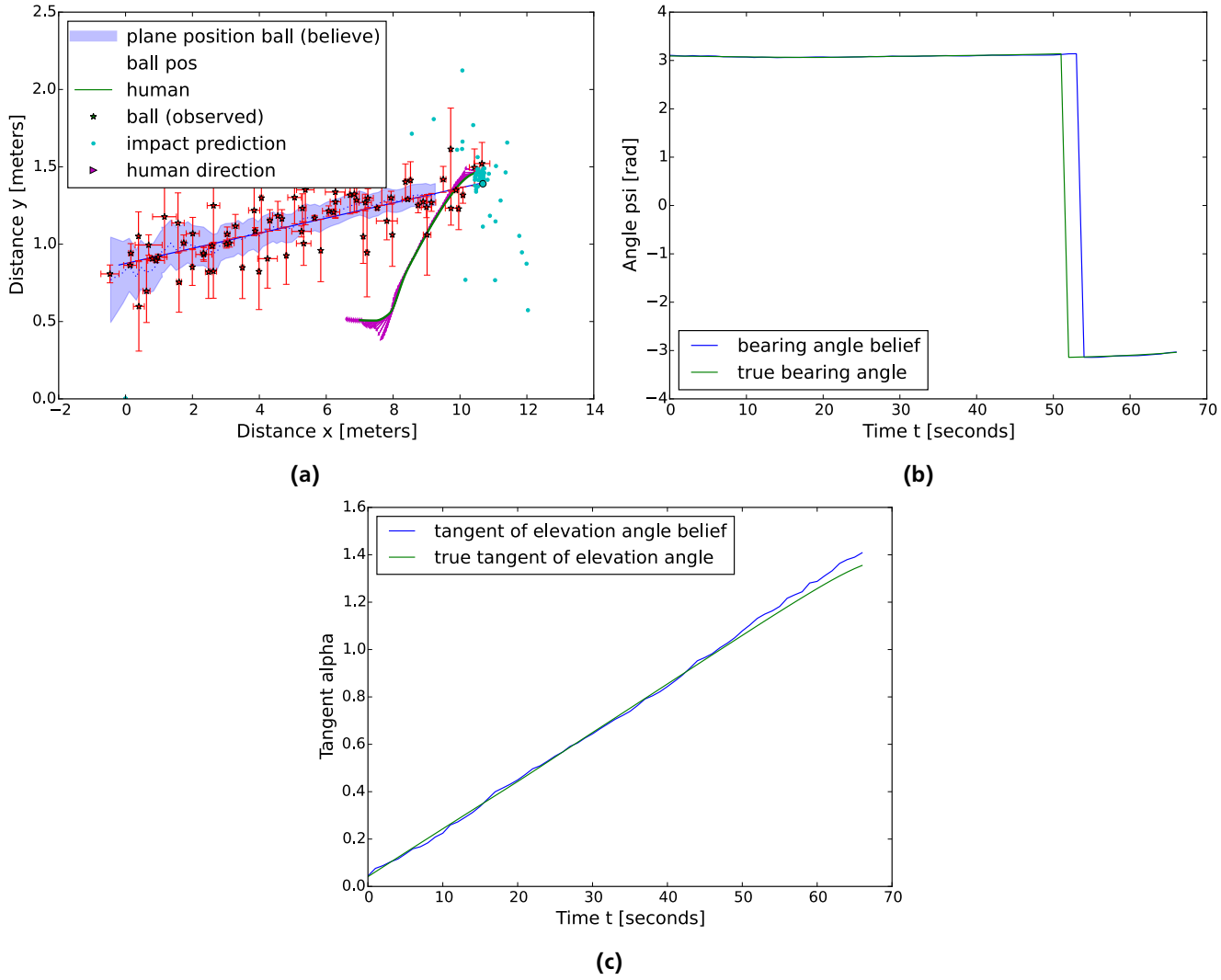


Figure E.4.: Example plots for a catch (the human is still closer than $20\,\mathrm{cm}$) with the simple policy. (a) This plot indicates the catch and shows that due to the backwards starting position (indicated by the backwards arrows) the human is always able to observe the ball. (b) The bearing angle stays constant as predicted by CBA. (c) The tangent of the elevation angle increases linearly as predicted by OAC.

The following is an example of a failed catch utilizing the best simple control policy found which shows typical results considering the usage of the heuristics for uncatchable balls.

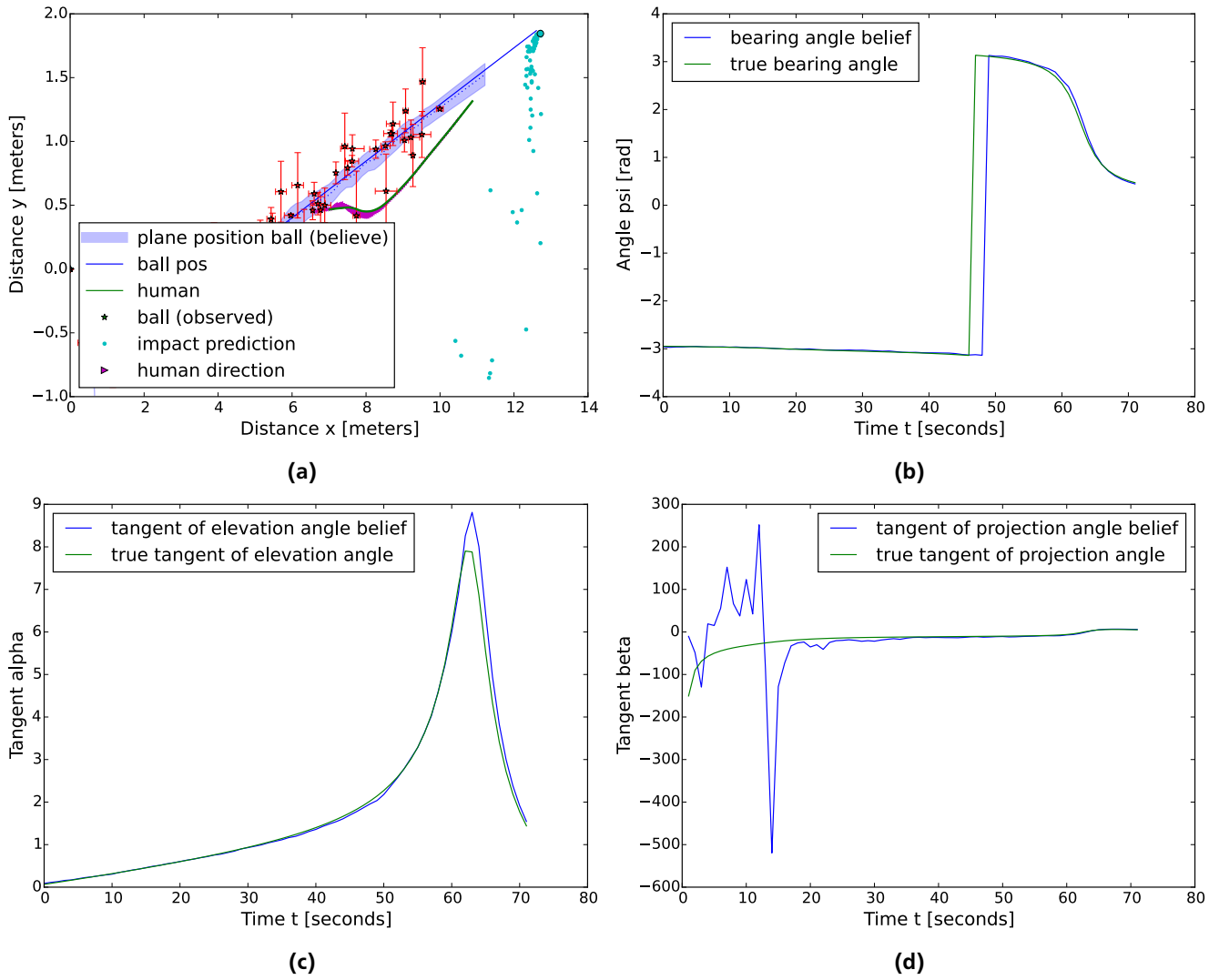


Figure E.5.: Example plots for a failed catch where the ball is simply too fast for the human to catch showing typical changes (in regard to the situation when the ball flies ahead of the human) in the tangent of the elevation angle and the bearing angle. (a) This plot indicates the failed catch and shows that due to the backwards starting position (indicated by the backwards arrows) the human is able to observe the ball until it flies over his head. (b) The bearing angle stays approximately constant as predicted by CBA until the ball flies ahead of the human at which point the bearing angle changes significantly. The same perturbation applies for OAC (c) but not for LOT. (d) The tangent of the projection angle is not constant.

The following plots show the switch from one bearing angle to another after the ball flies ahead of the human.

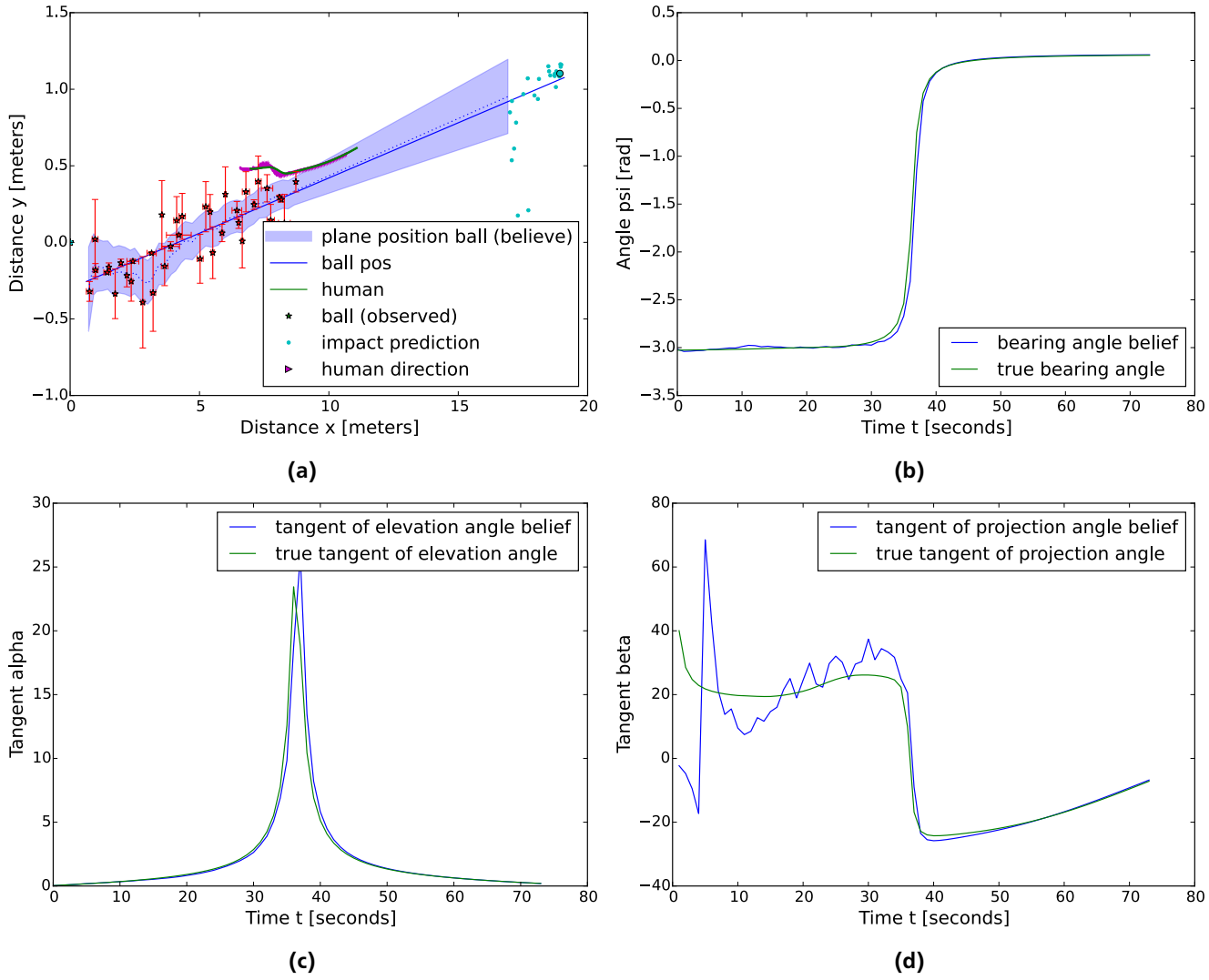


Figure E.6.: Example plots for a failed catch where the ball is too fast for the human showing typical changes (in regard to the situation when the ball flies ahead of the human) in the tangent of the elevation angle and the bearing angle. (a) This plot indicates the failed catch. (b) The bearing angle stays approximately constant as predicted by CBA until the ball flies ahead of the human at which point the bearing angle changes to a new one with a difference of $180°$. (c) The same perturbation applies for OAC which increases constantly at first and then decreases constantly as the ball flies ahead. (d) The tangent of the projection angle is not constant.

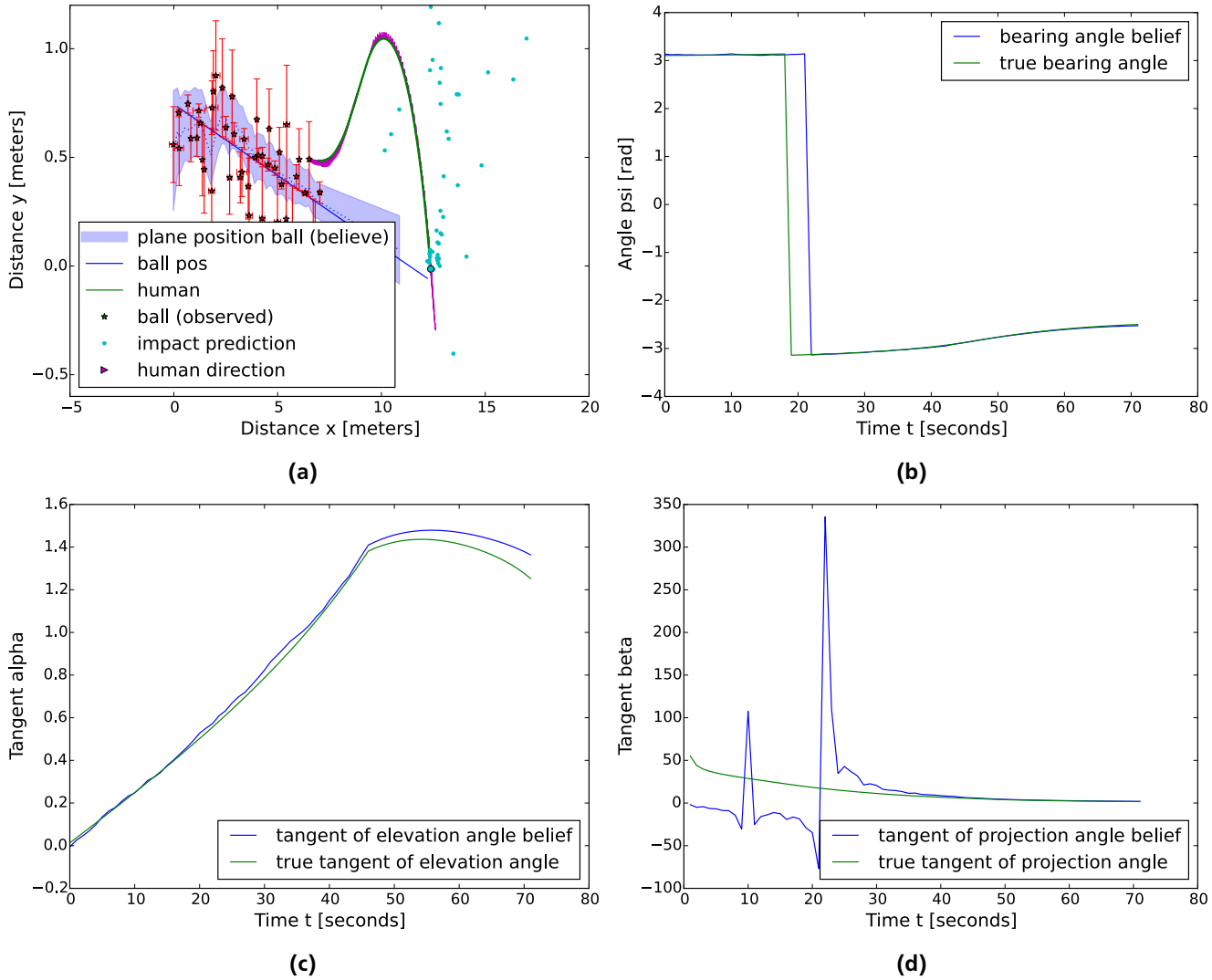


Figure E.7.: Example plots for a successful catch with the upper level policy. (a) This plot indicates the catch. (b) The bearing angle stays approximately constant. (c) The tangent of the elevation angle changes mostly linearly over time. (d) The tangent of the projection angle does not stay constant.

F Optimal Control Calculations

For a system

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t) + \xi_t$$

with $\xi_t \sim \mathcal{N}(0, \Sigma)$ Bellman's Principle of Optimality specifies a value function[1]

$$V(\mathbf{x}, \mathbf{u}) = \underset{\mathbf{u}}{\operatorname{argmax}}\, E\{V(f(\mathbf{x}, \mathbf{u}), t+1) + r(\mathbf{x}, \mathbf{u})\}$$

with

$$V(\mathbf{x}, T+1) = 0.$$

The system can also be written as

$$p(\mathbf{x}_{t+1}|\mathbf{x}, \mathbf{u}) \sim \mathcal{N}(\mathbf{x}_{t+1}|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{g}, \Sigma) = f(\mathbf{x}, \mathbf{u}) + \xi$$

using a Gaussian reward with the linear system as mean and a variance Σ.

With

$$E_{\mathbf{x}_{t+1}}\{g(\mathbf{x}_{t+1}|\mathbf{x}, \mathbf{u}\} = \int p(\mathbf{x}_{t+1}|\mathbf{x}, \mathbf{u}) g(\mathbf{x}_{t+1}) d\mathbf{x}_{t+1}$$

follows

$$\begin{aligned} V_T &= \underset{\mathbf{u}}{\operatorname{argmax}} \int p(\mathbf{x}_{t+1}|\mathbf{x}, \mathbf{u}) V(\mathbf{x}_{t+1}, T+1)) d\mathbf{x}_{t+1} + r_T(\mathbf{x}, \mathbf{u}) \\ &= \underset{\mathbf{u}}{\operatorname{argmax}}\, r_T(\mathbf{x}, \mathbf{u}) \\ &= \underset{\mathbf{u}}{\operatorname{argmax}} -\mathbf{x}^T\mathbf{R}\mathbf{x} - \mathbf{u}^T\mathbf{Q}\mathbf{u} \end{aligned}$$

which is maximal for $-\mathbf{u}^T\mathbf{Q}\mathbf{u} = 0$ and thus emerges a quadratic value function

$$V(\mathbf{x}, t) = (\mathbf{x} - \mathbf{v}_t)^T \mathbf{V}_t (\mathbf{x} - \mathbf{v}_t)$$

where $\mathbf{V}_t$ is a symmetric matrix and $\mathbf{v}_t$ a vector which represents any static values needed to reach the goal.

Now $\mathbf{V}_t$ and $\mathbf{v}_t$ need to be computed for every time step. For an easier computation the reward is transformed as follows (see Toussaint, 2009)

$$\begin{aligned} r(\mathbf{x}, \mathbf{u}) &= -(\mathbf{x} - \mathbf{v})^T \mathbf{R} (\mathbf{x} - \mathbf{v}) - \mathbf{u}^T\mathbf{Q}\mathbf{u} \\ &= -\mathbf{x}^T\mathbf{R}\mathbf{x} + 2\mathbf{v}^T\mathbf{R}\mathbf{x} - \mathbf{v}^T\mathbf{R}\mathbf{v} - \mathbf{u}^T\mathbf{Q}\mathbf{u} \\ r(\mathbf{x}, \mathbf{u}) &= -\mathbf{x}^T\mathbf{R}\mathbf{x} + 2\mathbf{r}^T\mathbf{x} - \mathbf{u}^T\mathbf{Q}\mathbf{u} \end{aligned}$$

$-\mathbf{v}^T\mathbf{R}\mathbf{v}$ can be ignored since it is constant and cannot change the reward.

Now a new value function is regarded which depends on t.

$$\begin{aligned} Q(\mathbf{x}, \mathbf{u}, t) = &\int \mathcal{N}(\mathbf{x}_{t+1}|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{g}, \Sigma)(-\mathbf{x}^T\mathbf{V}_{t+1}\mathbf{x} + 2\mathbf{r}^T\mathbf{x}) d\mathbf{x}_{t+1} - \mathbf{x}^T\mathbf{R}\mathbf{x} + 2\mathbf{r}^T\mathbf{x} - \mathbf{u}^T\mathbf{Q}\mathbf{u} \\ = &-(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{g})^T\mathbf{V}_{t+1}(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{g}) + 2\mathbf{r}^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{g}) + \operatorname{Tr}(\mathbf{V}_{t+1}\Sigma) \\ &- \mathbf{x}^T\mathbf{R}\mathbf{x} + 2\mathbf{r}^T\mathbf{x} - \mathbf{u}^T\mathbf{Q}\mathbf{u} \end{aligned}$$

[1] It would be $V(\mathbf{x}, \mathbf{u}) = \operatorname{argmax}_{\mathbf{u}}[V(f(\mathbf{x}, \mathbf{u}), t+1) + r(\mathbf{x}, \mathbf{u})]$ without considering noise

As the goal is to get the maximum for $\mathbf{u}$ the derivative $\dot{Q}(\mathbf{x},\mathbf{u},t)=0$ is calculated with respect to $\mathbf{u}$:

$$\begin{aligned}
-2(\mathbf{Ax}+\mathbf{Bu}+\mathbf{g})^T\mathbf{V}_{t+1}\mathbf{B}+2\mathbf{r}^T\mathbf{B}-2\mathbf{u}^T\mathbf{Q}&=0\\
-\mathbf{B}^T\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{Bu}+\mathbf{g})-\mathbf{B}^T\mathbf{r}-\mathbf{Q}^T\mathbf{u}&=0\\
-\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{Ax}-\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{Bu}-\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{g}-\mathbf{B}^T\mathbf{r}-\mathbf{Q}^T\mathbf{u}&=0\\
-\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{Bu}-\mathbf{Q}^T\mathbf{u}&=\mathbf{B}^T(\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{g})+\mathbf{r})\\
(-\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}-\mathbf{Q}^T)\mathbf{u}&=\mathbf{B}^T(\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{g})+\mathbf{r})
\end{aligned}$$

Rearranging the equation for $\mathbf{u}^*$ yields

$$\begin{aligned}
\mathbf{u}^* &= (-\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}-\mathbf{Q}^T)^{-1}\mathbf{B}^T(\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{g})+\mathbf{r})\\
&= -(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T(\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{g})+\mathbf{r})\\
&= \mathbf{K}_t\mathbf{x}+\mathbf{k}_t
\end{aligned}$$

since Q is a 1×1 matrix and with

$$\begin{aligned}
\mathbf{K}_t &= -\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{Ax}\\
\mathbf{k}_t &= -\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{g}-\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{r}
\end{aligned}$$

$$\begin{aligned}
Q(\mathbf{x},\mathbf{u},t) =& -\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}(\mathbf{Ax}+\mathbf{Bu}+\mathbf{g})-\mathbf{u}^T\mathbf{B}^T\mathbf{V}_{t+1}(\mathbf{Ax}+\mathbf{Bu}+\mathbf{g})\\
&-\mathbf{g}^T\mathbf{V}t+1(\mathbf{Ax}+\mathbf{Bu}+\mathbf{g})+2\mathbf{r}^T(\mathbf{Ax}+\mathbf{Bu}+\mathbf{g})+\mathrm{Tr}(\mathbf{V}_{t+1}\Sigma)\\
&-\mathbf{x}^T\mathbf{Rx}+2\mathbf{r}^T\mathbf{x}-\mathbf{u}^T\mathbf{Qu}\\
=& -\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{Ax}-\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{Bu}-\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{g}\\
&-\mathbf{u}^T\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{Ax}-\mathbf{u}^T\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{Bu}-\mathbf{u}^T\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{g}\\
&-\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{Ax}-\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{Bu}-\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{g}\\
&+2\mathbf{r}^T\mathbf{Ax}+2\mathbf{r}^T\mathbf{Bu}+2\mathbf{r}^T\mathbf{g}+\mathrm{Tr}(\mathbf{V}_{t+1}\Sigma)-\mathbf{x}^T\mathbf{Rx}+2\mathbf{r}^T\mathbf{x}-\mathbf{u}^T\mathbf{Qu}\\
=& -\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{Ax}-2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{Bu}-2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{Ax}-\mathbf{u}^T\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{Bu}\\
&-2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{Bu}-\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{g}\\
&+2\mathbf{r}^T\mathbf{Ax}+2\mathbf{r}^T\mathbf{Bu}+2\mathbf{r}^T\mathbf{g}+\mathrm{Tr}(\mathbf{V}_{t+1}\Sigma)-\mathbf{x}^T\mathbf{Rx}+2\mathbf{r}^T\mathbf{x}-\mathbf{u}^T\mathbf{Qu}\\
=& -\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{Ax}-\mathbf{x}^T\mathbf{Rx}\\
&-2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{Ax}+2\mathbf{r}^T\mathbf{Ax}+2\mathbf{r}^T\mathbf{x}\\
&-2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{Bu}-\mathbf{u}^T\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{Bu}-2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{Bu}+2\mathbf{r}^T\mathbf{Bu}-\mathbf{u}^T\mathbf{Qu}\\
&+\text{terms independent of }\mathbf{x}\\
=& -\mathbf{x}^T(\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{A}+\mathbf{R})\mathbf{x}-(2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{A}-2\mathbf{r}^T\mathbf{A}-2\mathbf{r}^T)\mathbf{x}\\
&-\mathbf{u}^T(\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{B}+\mathbf{Q})\mathbf{u}\\
&-2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{Bu}\\
&+(2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{B}-2\mathbf{r}^T\mathbf{B})\mathbf{u}+\text{terms independent of }\mathbf{x}\\
=& -\mathbf{x}^T(\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{A}+\mathbf{R})\mathbf{x}-(2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{A}-2\mathbf{r}^T\mathbf{A}-2\mathbf{r}^T)\mathbf{x}\\
&-[(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T(\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{g})+\mathbf{r})]^T\\
&\cdot(\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{B}+\mathbf{Q})(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T(\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{g})+\mathbf{r})\\
&+2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T(\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{g})+\mathbf{r})\\
&-(2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{B}-2\mathbf{r}^T\mathbf{B})(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T(\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{g})+\mathbf{r})\\
&+\text{terms independent of }\mathbf{x}\\
=& -\mathbf{x}^T(\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{A}+\mathbf{R})\mathbf{x}-(2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{A}-2\mathbf{r}^T\mathbf{A}-2\mathbf{r}^T)\mathbf{x}\\
&-(\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{g})+\mathbf{r})^T\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1^T}\mathbf{B}^T(\mathbf{V}_{t+1}^T(\mathbf{Ax}+\mathbf{g})+\mathbf{r})\\
&+2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T(\mathbf{V}_{t+1}^T\mathbf{Ax}+\mathbf{V}_{t+1}^T\mathbf{g}+\mathbf{r})\\
&-(2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{B}-2\mathbf{r}^T\mathbf{B})(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{Ax}+\text{terms independent of }\mathbf{x}
\end{aligned}$$

$$
\begin{aligned}
&= -\mathbf{x}^T(\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{A}+\mathbf{R})\mathbf{x} \\
&\quad -(2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{A}-2\mathbf{r}^T\mathbf{A}-2\mathbf{r}^T+(2\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{B}-2\mathbf{r}^T\mathbf{B})(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{A})\mathbf{x} \\
&\quad -((\mathbf{x}^T\mathbf{A}^T+\mathbf{g}^T)\mathbf{V}_{t+1}+\mathbf{r}^T)\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1^T}\mathbf{B}^T(\mathbf{V}_{t+1}^T\mathbf{A}\mathbf{x}+\mathbf{V}_{t+1}^T\mathbf{g}+\mathbf{r}) \\
&\quad +2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T(\mathbf{V}_{t+1}^T\mathbf{A}\mathbf{x}+\mathbf{V}_{t+1}^T\mathbf{g}+\mathbf{r}) \\
&\quad +\text{terms independent of } \mathbf{x} \\
&= -\mathbf{x}^T(\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{A}+\mathbf{R})\mathbf{x} \\
&\quad +2(-\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{A}+\mathbf{r}^T\mathbf{A}+\mathbf{r}^T-(\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{B}+\mathbf{r}^T\mathbf{B})(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{A})\mathbf{x} \\
&\quad -(\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}+\mathbf{g}^T\mathbf{V}_{t+1}+\mathbf{r}^T)\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1^T}\mathbf{B}^T(\mathbf{V}_{t+1}^T\mathbf{A}\mathbf{x}+\mathbf{V}_{t+1}^T\mathbf{g}+\mathbf{r}) \\
&\quad +2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T(\mathbf{V}_{t+1}^T\mathbf{A}\mathbf{x}+\mathbf{V}_{t+1}^T\mathbf{g}+\mathbf{r}) \\
&\quad +\text{terms independent of } \mathbf{x} \\
&= -\mathbf{x}^T(\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{A}+\mathbf{R})\mathbf{x} \\
&\quad +2(-(\mathbf{g}^T\mathbf{V}_{t+1}+\mathbf{r}^T)\mathbf{A}+\mathbf{r}^T-(\mathbf{g}^T\mathbf{V}_{t+1}+\mathbf{r}^T)\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{A})\mathbf{x} \\
&\quad -\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{B}+\mathbf{Q})^{-1^T}\mathbf{B}^T(\mathbf{V}_{t+1}^T\mathbf{A}\mathbf{x}+\mathbf{V}_{t+1}^T\mathbf{g}+\mathbf{r}) \\
&\quad -\mathbf{g}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{B}+\mathbf{Q})^{-1^T}\mathbf{B}^T(\mathbf{V}_{t+1}^T\mathbf{A}\mathbf{x}+\mathbf{V}_{t+1}^T\mathbf{g}+\mathbf{r}) \\
&\quad -r^T\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{B}+\mathbf{Q})^{-1^T}\mathbf{B}^T(\mathbf{V}_{t+1}^T\mathbf{A}\mathbf{x}+\mathbf{V}_{t+1}^T\mathbf{g}+\mathbf{r}) \\
&\quad +2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{A}\mathbf{x} \\
&\quad +2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{g} \\
&\quad +2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{r} \\
&\quad +\text{terms independent of } \mathbf{x} \\
&= -\mathbf{x}^T(\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{A}+\mathbf{R})\mathbf{x} \\
&\quad +2(-(\mathbf{g}^T\mathbf{V}_{t+1}+\mathbf{r}^T)\mathbf{A}+\mathbf{r}^T-(\mathbf{g}^T\mathbf{V}_{t+1}+\mathbf{r}^T)\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{A})\mathbf{x} \\
&\quad -\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}\mathbf{B}+\mathbf{Q})^{-1^T}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{A}\mathbf{x} \\
&\quad +2\mathbf{x}^T\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{A}\mathbf{x} \\
&\quad +\text{terms independent of } \mathbf{x} \\
&= -\mathbf{x}^T(\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{A}+\mathbf{R}-\mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{A})\mathbf{x} \\
&\quad +2(-(\mathbf{g}^T\mathbf{V}_{t+1}+\mathbf{r}^T)\mathbf{A}+\mathbf{r}^T-(\mathbf{g}^T\mathbf{V}_{t+1}+\mathbf{r}^T)\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{A})\mathbf{x} \\
&\quad +\text{terms independent of } \mathbf{x}
\end{aligned}
$$

Thus applies

$$
\begin{aligned}
\mathbf{V}_{t+1} &= \mathbf{A}^T\mathbf{V}_{t+1}\mathbf{A}+\mathbf{R}-\mathbf{K}\mathbf{V}_{t+1}^T\mathbf{A} \\
\mathbf{v}_{t+1} &= \mathbf{r}-\mathbf{A}^T(\mathbf{V}_{t+1}^T\mathbf{g}-\mathbf{r})-\mathbf{K}(\mathbf{V}_{t+1}^T\mathbf{g}-\mathbf{r})
\end{aligned}
$$

with

$$
\mathbf{K} = \mathbf{A}^T\mathbf{V}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{V}_{t+1}^T\mathbf{B}+\mathbf{Q})^{-1}\mathbf{B}^T
$$

Since the optimal value function is

$$
V^*(\mathbf{x},t) = V_t^*(\mathbf{x}) = \operatorname*{argmax}_u Q(\mathbf{x},\mathbf{u},t) = -\mathbf{x}^T\mathbf{V}_t\mathbf{x}+2\mathbf{x}^T\mathbf{v}_t + \text{terms independent of } \mathbf{x}
$$

for the optimal $\mathbf{u}$ as calculated above.

Bibliography

Robert K. Adair. *The Physics of Baseball*. New York: Harper and Row, 1990.

Jack B. Calderone and Mary K. Kaiser. Visual acceleration detection: Effect of sign and motion orientation. *Perception & Psychophysics*, 45(5):391–394, 9 1989.

Seville Chapman. Catching a baseball. *American Journal of Physics*, 36(10):868–870, 6 1968.

Ambreen Chohan, Martine H. G. Verheul, Paulien M. Van Kampen, Marline Wind, and Geert J. P. Savelsbergh. Children's use of the bearing angle in interceptive actions. *Journal of Motor Behavior*, 40(1):18–28, 2008.

Richard Dawkins. *The Selfish Gene*. Oxford University Press, 3 edition, 2006.

Gabriel Jacob Diaz, Flip Phillips, and Brett R. Fajen. Intercepting moving targets: a little foresight helps a lot. *Journal of Vision*, 195:345–360, 4 2009.

Brett R. Fajen and William H. Warren. Behavioral dynamics of intercepting a moving target. *test*, 180(2):303–319, 2 2007.

Philip W. Fink, Patrick S. Foo, and William H. Warren. Catching fly balls in virtual reality: A critical test of the outfielder problem. *Journal of Vision*, 9(13):1–8, 12 2009.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric numerical integration illustrated by the stoermer-verlet method. *Acta Numerica*, 12:399–450, 5 2003.

Nikolaus Hansen and Andreas Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. *In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation*, pages 312–317, 1996.

Verena Heidrich-Meisner and Christian Igeltest. Neuroevolution strategies for episodic reinforcement learning. *Journal of Algoritms Cognition, Informatics and Logic*, 64:152–168, 2009.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82:35–45, 3 1960.

Sadao Kawarnura, Fumio Miyazaki, and Suguru Arimototest. Is a local linear pd feedback control law effective for trajectory tracking of robot motion? *Robotics and Automation, 1988. Proceedings., 1988 IEEE International Conference on*, 3:1335 – 1340, 4 1988.

Julian N. Marewski, Wolfgang Gaissmaier, and Gerd Gigerenzer. Good judgments do not require complex cognition. *Springerlink.com*, 11:103–121, 9 2009.

Michael K. McBeath, Dennis M. Shaffer, and Mary K. Kaiser. How baseball outfielders determine where to run to catch fly balls. *Journal of Vision*, 268(5210):569–573, 4 1995.

Michael K. McBeath, Alan M. Nathan, A. Terry Bahill, and David G. Baldwin. Paradoxical pop-ups: Why are they difficult to catch? *American Journal of Physics*, 76:723–729, 2008.

Peter McLeod and Zoltan Dienes. Do fielders know where to go to catch the ball or only how to get there? *Journal of Experimental Psychology: Human Perception and Performance*, 22(3):531–543, 1996.

Peter McLeod, Nick Reed, and Zoltan Dienes. Toward a unified fielder theory: What we do not yet know about how people run to catch a ball. *Journal of Experimental Psychology: Human Perception and Performance*, 27(6):347–1355, 2001.

Javier R. Movellan. Primer on stochastic optimal control. *MPLab Tutorials, University of California San Diego*, 2009.

Bror V. H. Saxberg. Projected free-fall trajectories. *Biological Cybernetics*, 56:159–175, 5 1987.

John Schmerler. The visual perception of accelerated motion. *Perception*, 5:167–185, 2 1976.

Dennis M. Shaffer and Michael K. McBeath. Baseball outfielders maintain a linear optical trajectory when tracking uncatchable fly balls. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2):335–348, 7 2002.

Dennis M. Shaffer and Michael K. McBeath. Naive beliefs in baseball: Systematic distortion in perceived time of apex for fly balls. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6):1492–1501, 11 2005.

Dennis M. Shaffer, Scott M. Krauchunas, Marianna Eddy, and Michael K. McBeath. How dogs navigate to catch frisbees. *Psychological Science*, 15(7):437–441, 7 2004.

Thomas Sugar and Michael McBeath. Robotic modeling of mobile ball-catching as a tool for understanding biological interceptive behavior. *Behavioral and Brain Sciences*, 24(6):1078–1080, 12 2001a.

Thomas Sugar and Michael McBeath. Spatial navigation algorithms: Applications to mobile robotics. *Paper presented at the Proceedings of the 6th Vision Interface Annual Conference, Ottawa*, 6 2001b.

Thomas G. Sugar, Michael K. McBeath, Anthony Suluh, and Keshav Mundhra. Mobile robot interception using human navigational principles: Comparison of active versus passive tracking algorithms. *Autonomous Robots*, 21(1):43–54, 6 2006a.

Thomas G. Sugar, Michael K. McBeath, and Zheng Wang. A unified fielder theory for interception of moving objects either above or below the horizon. *Psychonomic Bulletin & Review*, 13(5):908–917, 10 2006b.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. The Mit Press, 2005.

Patrizio Tomei. Adaptive pd controller for robot manipulators. *Robotics and Automation, IEEE Transactions on*, 7(4):565 – 570, 8 1991.

Marc Toussaint. Robot trajectory optimization using approximate inference. *Proceedings of the 26th International Conference on Machine Learning*, pages 1049–1056, 2009.

Greg Welch and Gary Bishop. An introduction to the kalman filter. *TR*, pages 95–041, 7 2006.