1 Introduction

Many machine learning tasks involve high dimensional data in order to accurately represent their complex nature. The problem is that reinforcement learning is not scaling up well to accommodate the high dimensional data. For example, an usual approach for exploring a space with help of RL would be to count the exact states that have been visited already and through that calculate their individual relative probability. Now an even exploration can be achieved, when at each state the action is chosen that yields the lowest probability. The predicament that arises in high dimensional states is that they cannot be counted. In a real space the probability for each state would be zero and that can severely limit the scope of the exploration. So the dimensionality of the state itself has to be reduced to be able to work with the data
Two common methods to reduce the dimensionality of said data are the PCA and Deep Learning, i.e. autoencoders. In this thesis we want to look at both option and see how efficiently and evenly either approach explores the working space of a robot's endeffector. The aim of this thesis is, as mentioned previously, to find an optimal exploration strategy in continuous state-action space. Which means that preferably no single state occurs more than once. In the discrete case, We can measure that by counting how often a single state has been occupied and from that derive the state's probability of occurrence. Now whenever the exploration strategy needs to decide which action to do next, the available action, which reaches the state with the lowest probability, is chosen. But that is only practicable when the state is discrete. However, since the state of the robot is not in fact in discrete space – but in continuous space – we run into a problem. The probability that exactly this single state exists is infinitesimal. The result can be that the exploration will concentrate on a very tiny area within the search space, because even the smallest change of one of the state variables yields a probability of zero, for the state was not yet occupied. This is aggravated immensely by the increase of dimensions. The solutions to these kinds of problems are twofold. Firstly, not the probability of each single state is calculated but their combined density and furthermore, the dimensions of the state are being reduced to an amount that is more manageable to work with.

2 Preliminaries

This chapter covers the background knowledge that is required for this thesis. First the concept of reinforcement learning and its underlying method, the Markov Decision Process, is presented. The next section covers dimensionality reduction and the two reduction approaches, which are used in this thesis – PCA and an Autoencoder – are presented. The last section covers optimization problems and the CMA-ES, which is used throughout this thesis.

2.1 Markov Decision Process

The Markov Decision Process (MDP) is specified as a sequence of states and actions which adhere to the Markov property. It is mathematically defined as a five-tuple $M = (S, A, T, R, p_0)$, where

- S is the set of states
- A is the set of actions
- T is the transition probability function, such that $T(s, s', a) = p(s|s', a)$, where $s, s' \in S$ and $a \in A$
- R is the reward function, which assigns a reward r to every transition caused by action a
- p_0 is the initial distribution, which tells how likely it is to start in a certain state

2.2 Reinforcement Learning

Reinforcement learning is a field of machine learning, in which an agent shall learn a policy π to maximize the reward. The process is unsupervised, meaning there are no outside labels given and the solution is learned autonomously and from scratch. The environment in which the agent acts is usually formulated as a Markov Decision Process. The environment has a set of states S, the agent chooses from a set of actions A in each state and hence gets an intermediate reward r. The selection which action is chosen at which state is called policy and can be modeled as:

- $\pi : A \times S \mapsto [0, 1]$
- $\pi(s, a) = p(a|s)$

The aim now is to find a policy π, which maximizes the sum of all rewards. In value based reinforcement learning it is necessary to have knowledge of the accumulative reward starting at any state s. The value function V_π determines exactly that. It returns the expected accumulative reward when policy π was followed. It is defined as follows:

$$V^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s], \tag{1}$$

where $\gamma \in [0, 1]$, so that the effect of future states is counted less and less. It is called the discount factor. The maximum possible value of V^π is defined as:

$$V^*(s) = \max_\pi V^\pi(s) \tag{2}$$

When V^* has a solution, it is the best solution for the problem. However, it can be useful to also consider the action values. The function which does this, is defined as:

$$Q^{\pi}(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_{\mathrm{t}} | s_0 = s, a_0 = a] \tag{3}$$

The value $Q^{\pi}(s,a)$ returns the accumulative reward, starting at state s and choosing action a.

The goal now is to find a policy π, which tells the probability $\pi(s,a)$ that action a is chosen at state s. Under the condition that π^* is the optimal policy, optimality can be achieved, when at each state s the next action is chosen from $Q^{\pi^*}(s,\cdot)$.

2.3 Dimensionality Reduction

Dimensionality reduction is the process of transforming a given data set from a high-dimensional space into a low-dimensional space. This transformation has to be executed in such a way that the low dimensional representation of the data still retains the key properties of the original high-dimensional data for further process. There exist several ways of how to reduce the dimensions of a given data set. Two of them are discussed in detail in this thesis: Principal Component Analysis and dimensionality reduction through an autoencoder. Both these approaches are explained more thoroughly in the following chapters.

2.3.1 PCA

The Principial Component Analysis (PCA) is a method to reduce the dimensionality of a dataset by decorrelating its data and increasing its variance. It does this by transforming the data to so called principal components. The principal components are then ordered in a way, in which the first few of them that have the greatest eigenvalues hold the basic information of the original data set and account for the most variance.

The first step is to compute the covariance matrix of the entire dataset. To examine, whether there are any relations between different variables. The formula to calculate the covariance between to random variables X and Y is shown in ?:

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y)), \tag{4}$$

where n denotes the size of each random variable and $x_i \in X$ and $y_i \in Y$. $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ are referring to the means of X and Y, respectively. The covariance matrix for a 3-dimensional dataset would look like this:

$$\begin{pmatrix} Cov(X,X) & Cov(X,Y) & Cov(X,Z) \\ Cov(Y,X) & Cov(Y,Y) & Cov(Y,Z) \\ Cov(Z,X) & Cov(Z,Y) & Cov(Z,Z) \end{pmatrix}$$

Since the covariance of a variable with itself is its variance, the diagonal of the matrix consists of the variance of each random variable. And since furthermore, the covariance is commutative, the matrix is mirrored at the diagonal.

The second step in the PCA is to calculate the eigenvectors and eigenvalues of the covariance matrix. This is done to ascertain the principal components. The formula for calculating the eigenvalues of a matrix is shown in ?:

$$det(\mathbf{A} - \lambda\mathbf{I}) = 0, \tag{5}$$

where $\mathbf{A}$ denotes the matrix, $\mathbf{I}$ the identity matrix and λ the eigenvalues to be calculated.
So in the case of the 3-dimensional covariance matrix, the calculation is shown below:

$$Cov \begin{pmatrix} Cov(X,X) - \lambda & Cov(X,Y) & Cov(X,Z) \\ Cov(Y,X) & Cov(Y,Y) - \lambda & Cov(Y,Z) \\ Cov(Z,X) & Cov(Z,Y) & Cov(Z,Z) - \lambda \end{pmatrix} = 0$$

For each eigenvalue λ there is a corresponding eigenvector v. They can be calculated with the formula shown in ?:

$$(\mathbf{A} - \lambda_i\mathbf{I})v_i = 0 \tag{6}$$

where $i = 1, ..., m$ and m denotes the number of eigenvalues.

Since the goal is to reduce the dimensionality of the dataset, the next step is to sort the eigenvectors (principal components) and only keep the k of them, which account for the most variance. Because the eigenvectors simply indicate the direction of the new axes, the sorting is done based on the corresponding eigenvalues. So the k eigenvectors are chosen that possess the highest eigenvalues. These are then used to form a new d $\times$ k dimensional matrix $\mathbf{W}$.

Now the last step is to multiply this matrix $\mathbf{W}$ with the original dataset to project it into the lower dimensional subspace. The formula for this is shown in ?:

$$\mathbf{S} = \mathbf{O} \times \mathbf{W}, \tag{7}$$

where $\mathbf{S}$ denotes the projection in the subspace and $\mathbf{O}$ the original dataset.

Sometimes the last step is to also whiten the data to make it even less correlated and to give all data points the same variance. This is done by dividing every dimension (column) of the matrix $\mathbf{W}$ by the square root of its corresponding eigenvalue.

2.3.2 Autoencoder

The special kind of neural networks used in this thesis is called autoencoder and they are wildly used in unsupervised learning. Their purpose is to reduce the dimensionality of a data set while keeping its key features. They differ from PCA in the way they transform the data. While PCA performs only linear transformations, autoencoders perform nonlinear transformation, where the nonlinearity emerges from the noninear activation in the neural network.

An autoencoder is usually comprised of two main parts: The encoding part and the decoding part. Since autoencoders are usually symmetric in nature, the decoder has the same amount of layers as the encoder. It incorporates the hidden layer(s) and the output layer (see figure ?).

The idea is that the encoder subsequently reduces the dimensionality of the input through several different layers to a compressed representation of the input. Following that the decoder increases the dimensionality of the now compressed input until the input and the output have the same dimension. Given that procedure, an autoencoder is classified as an unsupervised learning model.

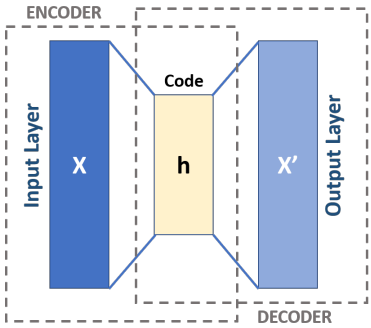


Figure 1: Simple Autoencoder: Placeholder

The two parts of an autoencoder can be mathematically represented as follows:

- The encoder is a function $f : X \mapsto H$ that maps the n-dimensional input array $X \in \mathbb{R}^n$ to the compressed representation $H \in \mathbb{R}^m$, such that $m < n$.
- The decoder is a function $g : H \mapsto Y$ that maps the m-dimensional compressed representation $H \in \mathbb{R}^m$ to the reconstructed data $Y \in \mathbb{R}^n$

The autoencoder itself can be represented as the concatenation of these two functions:

- Choose f and g, such that $f, g = \underset{f,g}{\operatorname{argmin}} |X - g(f(X))|$.

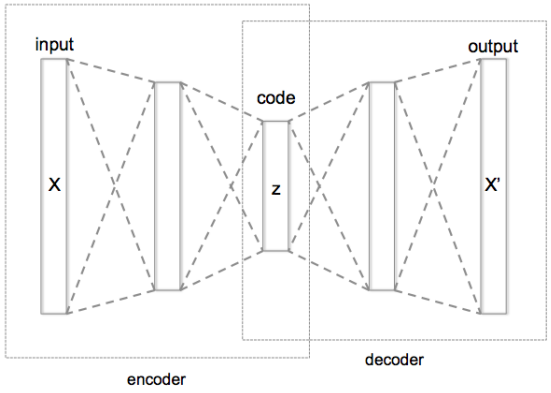


Figure 2: Simple Autoencoder: Platzhalter

2.4 Density Estimation

Density estimation is a way to estimate the probability density function of given data. The probability density function describes the relative likelihood that a specific sample (here one dimension of the reduced state-action-pair) occurs anywhere in the search radius. An example of that is the normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$, where μ is the mean and σ^2 the variance. The probability that a state occurs that is less then one σ away from μ is 68.26 percent. Now the exploration strategy does not have to choose the next state with the lowest probability, but with the lowest probability density. However, this estimation has to be computed for every dimension of the state.

2.5 Optimization Problems

2.5.1 Definition of an Optimization Problem

An optimization problem is the problem of finding the best solution from all available solutions. It can be mathematically represented as shown in ?:

$$\begin{aligned} x^* &= \underset{x}{argmin} f_0(x) \\ \text{subject to } & f_\text{i}(x^*) \leq b_\text{i},\ i = 1, , , m. \end{aligned} \tag{8}$$

Here $f_0 : \mathbb{R}^\text{n} \mapsto \mathbb{R}$ is called the objective function and $f_\text{i} : \mathbb{R}^\text{n} \mapsto \mathbb{R}$, $i = 1, ...m$ the constraint function. $b_1, ..., b_\text{m}$ refer to the constraints or limits. A vector x_* is then optimal, when it has the smallest objective value (see ?.1), while satisfying all constraints (see ?.2).

2.5.2 Evolution Strategy

An evolution strategy is a technique that is primarily used for optimization problems. As the name suggests, this technique tries to find the optimal solution for a problem through the methods of evolution: Mutation and Selection. Several implementations exist that differ in what they mutate and what they select. But the general process is, it generates a set of candidate solution which it analyzes on the basis of a fitness or an objective function. The proposed solutions that yield the best fitness values are then used to generate the next generation of candidate solutions. This process only ceases until a predefined criteria has been met.

2.5.3 The CMA-ES

One kind of evolution strategies proposes new candidate solutions by randomly sampling from a multivariate normal distributions with mean μ and a fixed covariance matrix $\mathbf{\Sigma}$. In each generation the current mean is updated based on the best candidates from the previous generation. However, because the covariance matrix $\mathbf{\Sigma}$ is fixed and with that the search radius, one shortcoming is that whenever $\mathbf{\Sigma}$ is inadequately chosen, the convergence speed can be slow for $\mathbf{\Sigma}$ is too small or even worse, the optimization process gets stuck in a local optimum.

The **C**ovariance **M**atrix **A**daption **E**volution **S**trategy (CMA-ES) is a special kind of an evolution strategy that overcomes this issue, since it updates not only the mean μ every generation but also the covariance matrix $\mathbf{\Sigma}$. As is illustrated in figure ?, this modification allows for a large search space in the beginning and thus a fast convergence to the optimum and a smaller search space towards the end for finetuning the found optimum.

The general procedure of the CMA-ES can be presented as follows:

- Create multivariate normal distribution $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ (The initial values are usually $\mu_0 = 0$ and $\mathbf{\Sigma}_0 = I$)
- Sample N points from X, such that $Y = (y_1, ..., y_N)$ with $y_i \in X \; \forall \; i = 1...N$
- Evaluate all samples from Y with a previously defined fitness function f, such that $F = (f(y_1), ..., f(y_N)) \; \forall \; y_i \in Y$
- From F choose the M samples with the best fitness value (i.e. the highest or the lowest) and calculate the new mean μ and the new covariance matrix $\mathbf{\Sigma}$

This procedure is repeated until a termination criteria has been met, for example a certain amount of generations have passed or a certain threshold is surpassed.

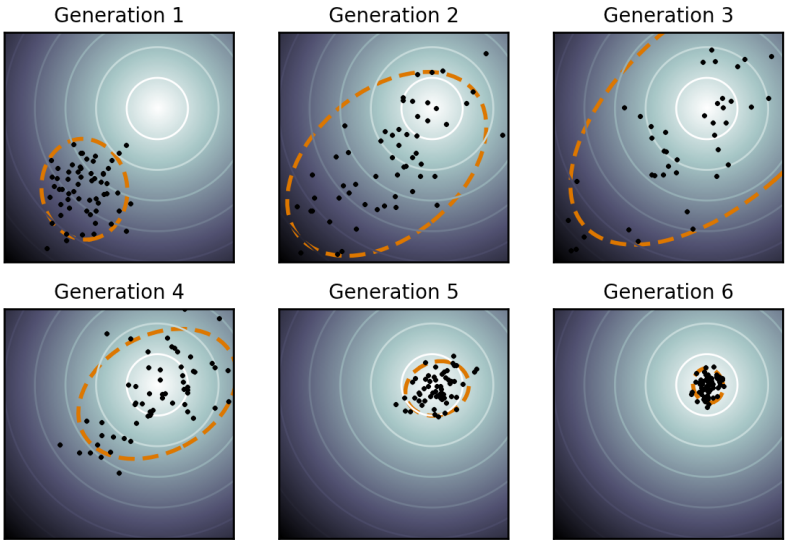


Figure 3: CMA-ES: Placeholder

3 Related Works

3.1 Exploration strategy in continuous state-action space

This thesis proposes a novel exploration strategy for continuous state-action space based on state novelty.

There are however several other approaches to achieve exploration in a continuous state-action space. These can be grouped in three main categories: Prediction error, information gain and state novelty.

3.1.1 Exploration through Prediction Error

The methods, which belong to the category of exploration through prediction error, generate an intrinsic reward based on the difficulties an agent has to predict the outcomes of its actions. This however can be lead to many difficulties given the stochastic nature of the agent's environment and the agent's actuators inherent noise.
Deepak Pathak et al (2017) [cit] managed to avoid many problems with previous prediction approaches due to only predicting the changes in the environment of the agent that could only be due to its actions. They did not let the agent make predictions based on the raw input from its sensors, but they transformed the input into a feature vector that contained only the information relevant to the agent's actions. This feature space is learned by a neural network. Then a forward dynamics model is trained from the feature space, which is used to

predict the feature representation of the next state, given the current state and action. The prediction error serves as the intrinsic reward. They observed that the agent manages to move around the corridors and halls of the game *VidDoom* without any extrinsic reward. In the game *Mario* however, the agent only manages to complete around 30% of the first level due to a specific sequence of button presses that the agent has to do. If it does not, it is not able to explore the level further. They concluded that their methods performs well when the environment has only few reward signals, but does not extend to environments that have few interaction opportunities as well.

3.1.2 Exploration through Information Gain

The methods, which belong to the category of exploration through information gain, generate an intrinsic reward based on the reduction of entropy that the action can bring.

3.1.3 Exploration through State Novelty

The methods, which belong to the category of exploration through state novelty, generate an intrinsic reward on the novelty of the state that an action brings. In discrete state spaces this can be easily done by keeping records of the already explored states and thus knowing the probability of each state. Unfortunately this approach can not be extended to continuous state spaces for the probability of any state would always be zero .
Bellemare et al (2016) [] proposed pseudo-counts that were directly derived from the density model. Based on this pseudo-count, they generated an exploration bonus for a Deep-Q-Network-agent. This agent managed to achieve state of the art in the game *Montezuma's Revenge* on the Atari 2600 when combined with a mixed Monte Carlo update. Ostrovski et al (2017) [] examined how important the quality of the density model is and and what role the Monte Carlo update plays. They compared a pseudo-count derived from the CTS density model and a pseudo-count derived from the PixelCNN density model. They found that the PixelCNN density model produces somewhat better samples. FOr the role of the Monte Carlo update they found that it speeds up training but hurts performance.

3.2 Smooth Trajectories

Since this thesis is focusing on finding an exploration strategy for robots, it has to be guaranteed that the new action, which is chosen by the exploration strategy (in this case a new velocity), does not harm the actuators of the robot. This is why it is necessary to achieve a smooth transition from the old action to the new. That is achieved by representing the robot's action as a movement primitives.

4 Proposed Methods

4.1 Interpolating trajectory

In reinforcement learning a new action is chosen at each decision step. This is true for discrete and continuous state spaces. The problem, which arises is that this approach does not consider a smooth trajectory, i.e. what happens in between the decision steps. Furthermore, in a real robot task the hardware oscillates as well. One potential solution for this is to only allows a small action amplitude. However, they are also non smooth. Therefore, in this thesis, another approach, which is examined is based on the idea of trajectory blending in movement primitives.

In this thesis a new actions is chosen every second. This action is a vector of velocities of each joint. Within the simulation environment an abrupt change of velocity is not a problem. However, in real robots this can severely harm its mechanical parts. Therefore it is necessary to enable a smooth transition from the current action to the new action, i.e. from the current to the new velocity. Since the simulation works in 50 ms steps and since a new action is chosen every second, 20 points between two defined actions have to be interpolated. The formula, which was uses to achieve this is shown in ?. Its derivative is shown in ?:

$$V_{\mathrm{i}}(t) = (1 - \sin(t * \frac{\pi}{2})^3) * V_{\mathrm{c}} + \sin(t * \frac{\pi}{2})^3 * V_{\mathrm{n}} \tag{9}$$

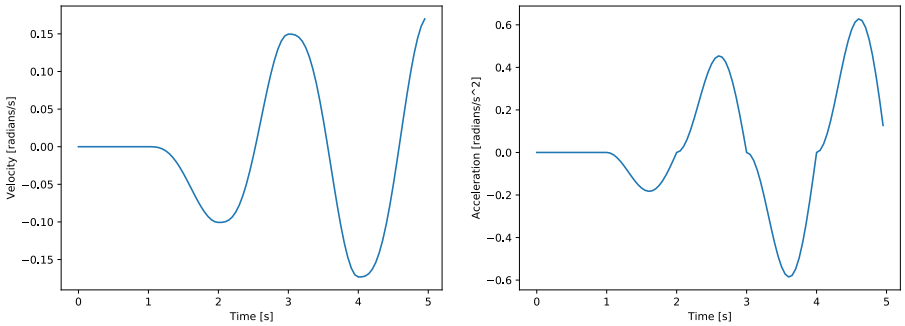
$$A_{\mathrm{i}}(t) = -\frac{3}{2}\pi \cos(t * \frac{\pi}{2}) \sin(t * \frac{\pi}{2})^2 (V_{\mathrm{c}} - V_{\mathrm{n}}) \tag{10}$$

Here V_i and A_i denote the current interpolated velocity and acceleration, respectively. V_c denotes the current velocity and V_nt the new velocity. t stands for the time that has passed within the one second interval in 50 millisecond steps, i.e. 50ms, 400ms or 850ms. The velocity profile is shown in figure ?.1 and the acceleration profile in figure ?.2, respectively.

Figure ? shows the velocity and the acceleration profile for five iterations. The velocity starts at 0 Radians/s (time step 1, figure ?.1). The following actions are: -0.1 Radians/s (time step 2), 0.1497 Radians/s (time step 3)and -0.1732 Radians/s (time step 4). This sequence of actions leads to the velocity plot shown in figure ?.1. The trajectory between the starting and the end velocity of each decision step (t = 0,1,2 etc.) is smoothly blended.
The acceleration profile for this sequence of actions is shown in figure ?.2. The acceleration at each decision step is zero and the trajectory in between is smooth.

4.2 Selection of new Actions

This thesis presents three different approaches for the exploration of a continuous state-action space. Each of these three approaches uses a different method of how to acquire a new set of actions. These methods are stochasticity, autoencoder and PCA and will be covered in the following chapters.



(a) Velocity profile

(b) Acceleration profile

Figure 4: Plots of the velocity and acceleration profile.

4.2.1 Exploration by Stochasticity

The first method of acquiring a new set of actions involves sampling from a normal distribution $X \sim \mathcal{N}(\mu, \sigma)$. This is the simplest method computation wise and also the quickest overall. A random value is drawn from the normal distribution, where the mean μ is equal to zero and the variance σ is equal to $\frac{constraint}{\sqrt{2}}$ (see figure ?, line 11). In this thesis the constraint is referring to the angular speed of the joint and is set to ten degrees. This process is repeated for every joint, seven in this thesis (see figure ?, line 4).

After an action is sampled for each joint, it is determined whether the actions satisfy all the constraints. If even one of them is violated, the entire process is started anew (see figure ?, line 10).

Algorithm 1 Base Line

```
1:  procedure MAIN
2:      EPISODES ← N
3:      ITERATIONS ← M
4:      K ← Number of joints
5:      constraint ← Max joint velocity
6:      Start environment
7:
8:      for i ← 1, EPISODES do
9:          for j ← 1, ITERATIONS do
10:             while constraint is not satisfied do
11:                 Sample new set of actions {a_{k,j}}_{k=1}^K with N(0, constraint/√2)
12:             Interpolate trajectory between {a_{k,j-1}}_{k=1}^K and {a_{k,j}}_{k=1}^K
13:             Reset environment
14:
```

Figure 5: Pseudocode Baseline

4.2.2 Exploration through an Autoencoder and the CMA-ES

Another way of acquiring a new set of actions is through the CMA-ES and an autoencoder. It has already been explained how they work and what they are used for individually (see chapter ? and ?), but now we use them in conjunction. The process itself also uses density estimation (chapter ?) an evaluation method for reinforcement learning in a continuous state-action space.

The first part is to set up the autoencoder itself. In this thesis a two layer approach has been chosen. The first layer reduces the 21 dimensional input vector to a 16 dimensional vector. This vector is reduced further to five dimensions in the second layer.
For the activation function PReLU [citation needed] was used. After the autoencoder has been set up, it is trained after each episode with all the to this point acquired data for twenty epochs (see figure ?,line 17).

The second step is the actual sampling of the set of actions with the help of the CMA-ES. In each iteration of the program the CMA-ES is initialized with the last set of actions (the current velocity), a standard deviation of 0.25, a population size of 16 and a maximum amount of iterations of 50 (see figure ?, line 10). Consequently the CMA-ES runs at most 50 iterations and in each iteration 16 sets of candidate solutions are being sampled by the algorithm. Each of those sets gets concatenated with the current state and propagated through the previously trained autoencoder to reduce their dimensionality and obtain their compressed representation. This compressed representation of the current state-actions-pairs is then evaluated on the compressed representation of the state-actions-pairs of the last N episodes. [1] The set of actions that yields the best result (lowest density) is then chosen as the base for the next iteration of the CMA-ES. This is repeated until no better result can be obtained or 50 iterations are reached. The best overall result is then chosen as the new action for this iteration of the simulation (see figure ?, line 11).

4.2.3 Exploration through PCA and CMA-ES

In the third approach for acquiring a new set of actions, the set of actions itself is also determined by the CMA-ES. However, the method of reducing the dimensionality of the state-action pair differs from the autoencoder approach. Here this is achieved through PCA.

In each iteration the data of the last N (20 in this thesis) episodes is compressed in its dimensionality using PCA (see figure ?, line 14). Then a new set of actions is acquired. The process of how this is done exactly is the same as in the previous approach. The sets of actions that are proposed by the CMA-ES

[1] N can be chosen as one likes. In this thesis N is fixed to 20, so the overall calculation time could be reduced. However, N can also be set to incorporate every previous episode

Algorithm 2 Autoencoder

procedure MAIN
  $EPISODES \leftarrow N$
  $ITERATIONS \leftarrow M$
  $K \leftarrow$ *Number of joints*
  constraint $\leftarrow$ *Max joint velocity*
  Initialize Autoencoder

  for $i \leftarrow 1$, *EPISODES* **do**
    for $j \leftarrow 1$, *ITERATIONS* **do**
      while *constraint not satisfied & max iteration not reached* **do**
        ▷ X refers to set of candidate solution proposed by CMA-ES
        $\{a_{k,j}\}_{k=1}^{K} \leftarrow \underset{x \in X}{\text{argmin}}\, \text{density}(AE(\{s_{k,j}, a_{k,x}\}_{k=1}^{K}))$
      Interpolate trajectory between $\{a_{k,j-1}\}_{k=1}^{K}$ and $\{a_{k,j}\}_{k=1}^{K}$
      Add current state and action $\{s_{k,j}, a_{k,j}\}_{k=1}^{K}$ to replay buffer R
      Reset environment

    Train Autoencoder
    Save hidden layer data

Figure 6: Pseudocode Autoencoder

are also compressed using the same method. The density estimation of this compressed representation of the sets of actions is then evaluated on the density estimation of the compressed representation of the data of the last N episodes. The set of actions that yields the best result (lowest density) is then chosen as the base for the next iteration of the CMA-ES. This is then repeated, as it was in the autoencoder approach as well, until the best result has been found or 50 iterations have been surpassed (see figure ?, line 9f).

Algorithm 3 PCA

procedure MAIN
  $EPISODES \leftarrow N$
  $ITERATIONS \leftarrow M$
  $K \leftarrow$ *Number of joints*
  constraint $\leftarrow$ *Max joint velocity*

  for $i \leftarrow 1$, *EPISODES* **do**
    for $j \leftarrow 1$, *ITERATIONS* **do**
      while *constraint not satisfied & max iteration not reached* **do**
        ▷ X refers to set of candidate solution proposed by CMA-ES
        $\{a_{k,j}\}_{k=1}^{K} \leftarrow \underset{x \in X}{\text{argmin}}\, \text{density}(AE(\{s_{k,j}, a_{k,x}\}_{k=1}^{K}))$
      Interpolate trajectory between $\{a_{k,j-1}\}_{k=1}^{K}$ and $\{a_{k,j}\}_{k=1}^{K}$
      Add current state and action $\{s_k, a_{k,j}\}_{k=1}^{K}$ to replay buffer R
      Fit PCA with state-action-pairs from replay buffer R
      Reset environment

Figure 7: Pseudocode PCA

5 Experimental Setup

5.1 Environment Setup

The robot that was used throughout this thesis is the kuka lbr iiwa 7 r800 fabricated by the company called KUKA AG. To simulate the behavior of said robot, the software CoppeliaSim by Coppelia Robotics was used. The code itself is written in Python.

5.2 Task formulation as an RL problem

A state is defined as a 14-dimensional array. The first seven entries are the joint positions of the robot measured in Radian. The last seven values are the current velocity of each joint of the robot measured in Raidan per second. The action is a seven dimensional array, which holds the new joint velocities that the robot should reach. Together both state and action form the state-action-pair, which is mentioned throughout the thesis.

One second in the simulation consists of 20 50 millisecond steps. In this thesis this is referred to as one iteration. At the start of the iteration a new action is chosen and at the end of the iteration the state velocity should be equal to the action.

100 iteration make up one episode. The different approaches were tested with a different number of episodes. The baseline and the PCA approach both ran for 200 episodes and the auroencoder approach ran for 500 episodes.

6 Results

For the presentation two kinds of plots are available. Firstly the plot that shows all the positions where a specific joint has been during all iterations and all episodes in Cartesian (x,y,z) coordinates. The second shows a density estimation of the position of a joint in radian. It illustrates how often a specific angle has been occupied.
The autoencoder has an additional kind of plots. It shows the density estimation of the compressed representation of the input layer.

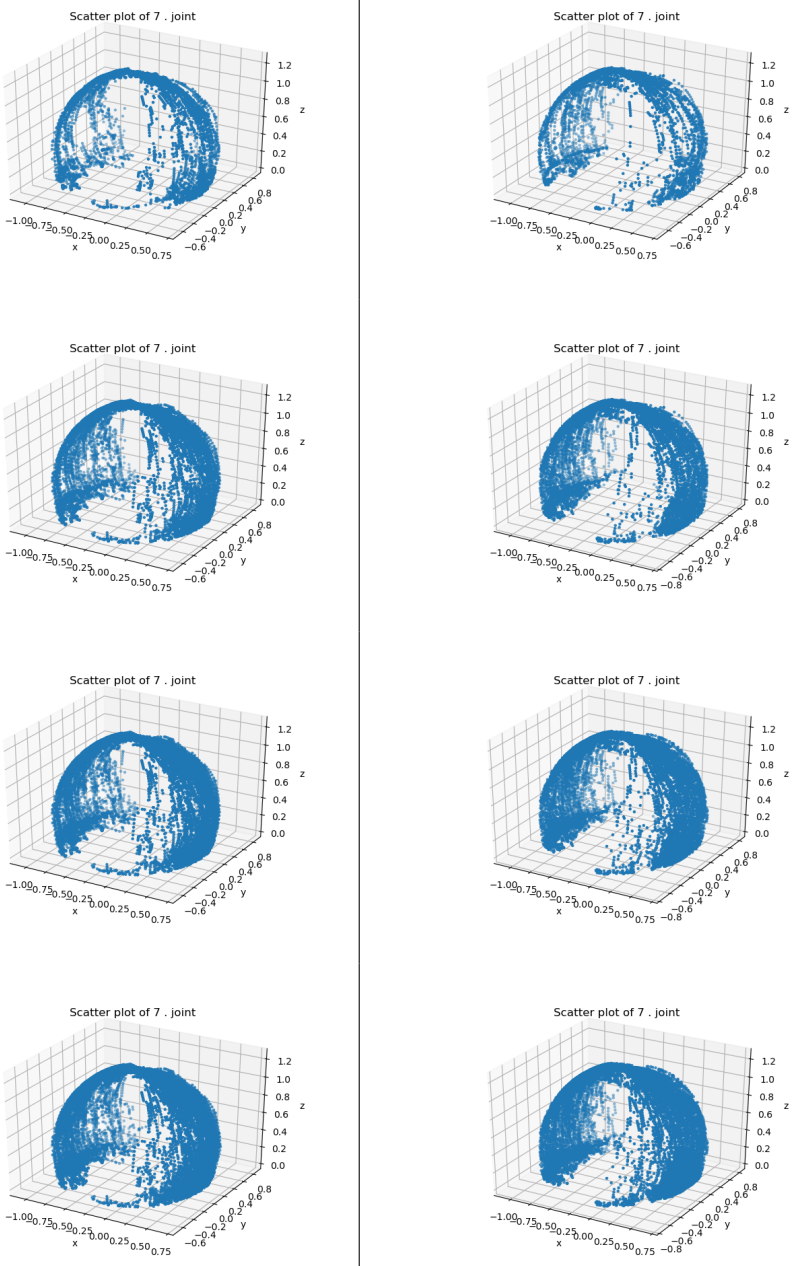
Only a part of the the research is shown here directly. Other plots can be found in the appendix.

6.1 Baseline Exploration

First we want to take a look on how the baseline exploration fared in exploring the working space of the robot. As a quick reminder, the next action in this approach was randomly sampled from a normal distribution. In figure ? we can see the result of that sampling method. The plots show all the positions the

endeffector has occupied after 60, 120, 160 and 200 episodes for two examples in Cartesian Coordinates. We can see that only a small portion of the entire working space has been explored by the robot's endeffector. If we also compare the progression of the exploration after certain episodes, we can also observe, than the explored room itself is only shrinking slowly, since the endeffector keeps visiting the same coordinates.

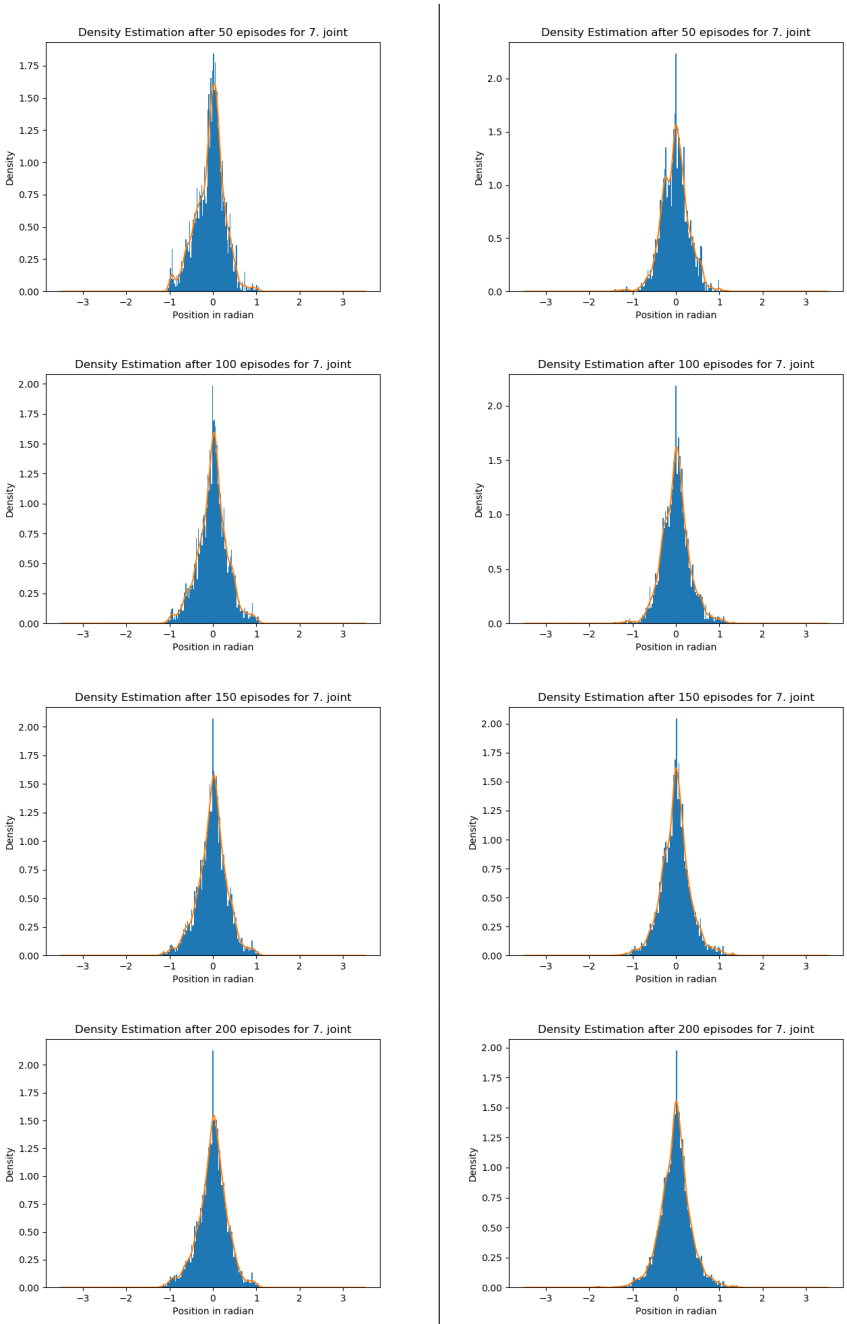
The effect of the sampling through the normal distribution can be seen in figure ? as well. It shows plots of the density of the robot's inherent endeffector position in Radians for the same two examples as in figure ?. As wee can see, the density looks similar to a normal distribution. Most of the density is centered around 0 degree (the mean of the normal distribution). Additionally we can see that the density quickly fades to zero and never reaches the constraint that the joint has (3.054326 Radian in this case). The same can be observed for all other joints as well (see Appendix).



Plot of endeffector of one simulation after 60, 120, 160 and 200 episodes

Plot of endeffector of another simulation after 60, 120, 160 and 200 episodes

Figure 8: Comparison of the endeffector positions after different episodes



Plot of density of endeffector position of one simulation after 60, 120, 160 and 200 episodes

Plot of density of endeffector position of another simulation after 60, 120, 160 and 200 episodes

Figure 9: Comparison of the endeffector position density after different episodes

6.2 Autoencoder exploration

6.2.1 Training of the autoencoder

At the end of each episode the autoencoder is trained with all the to this point acquired state-action-pairs. What setting were used, you can see in the ??? chapter. From figure ? we can see that the loss is roughly between 0.05 and 0.35.

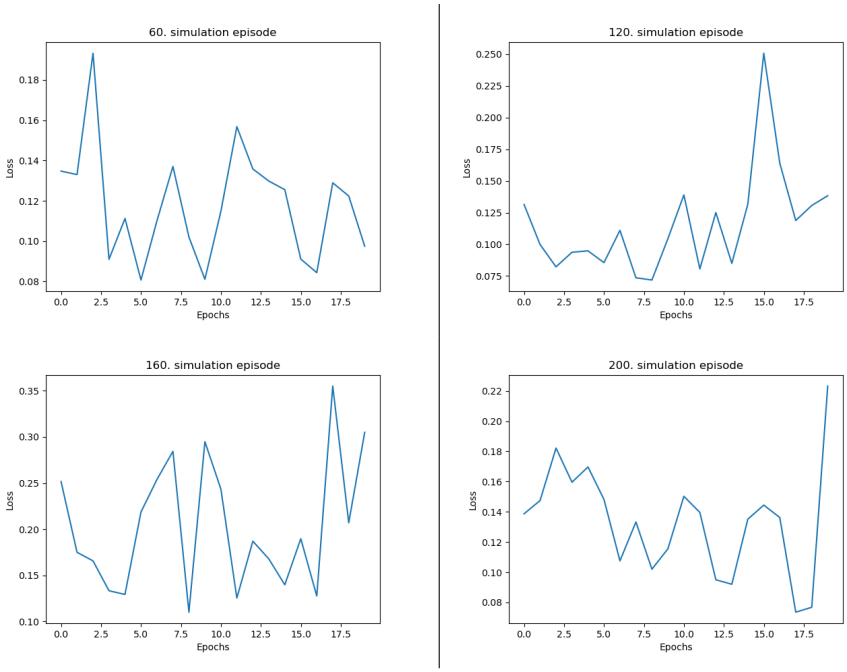


Figure 10: Comparison of the loss of the autoencoder after 60, 120, 160 and 200 episodes

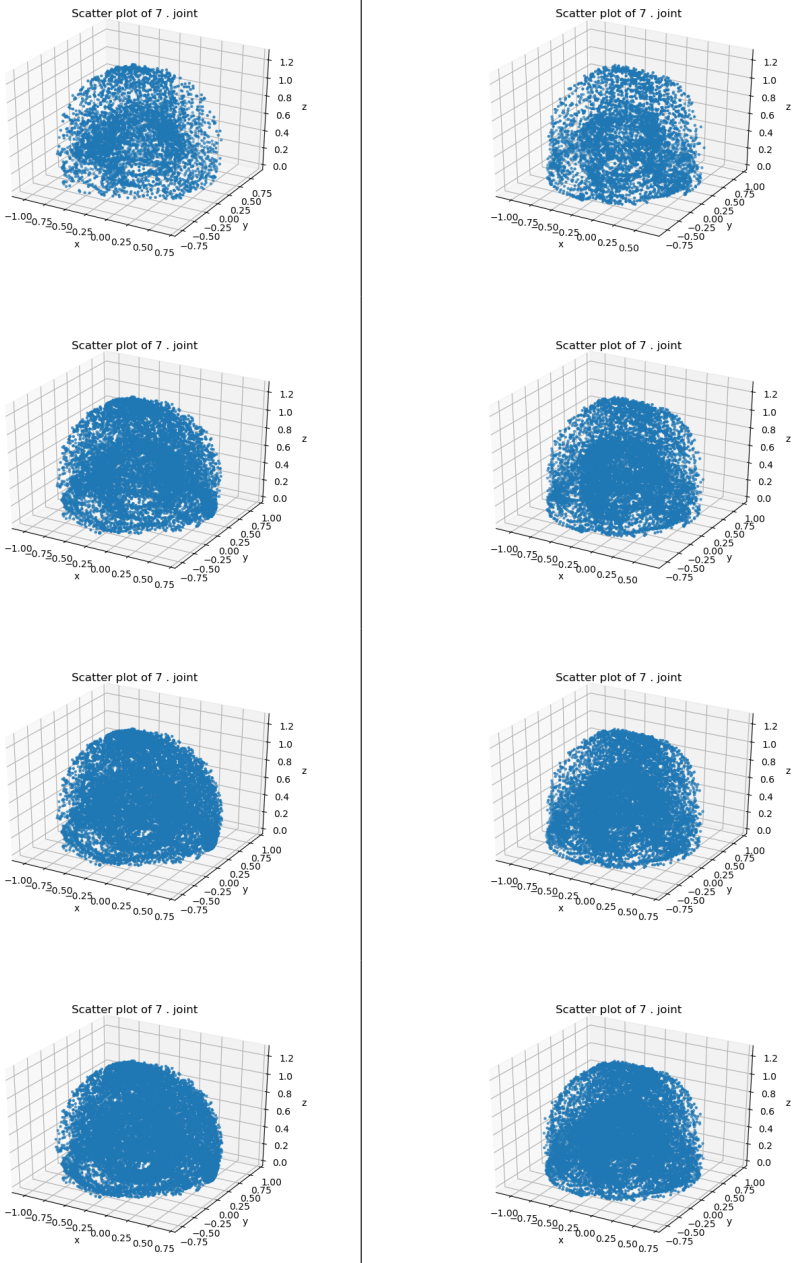
6.2.2 Exploration of the working space

Now we take a look how the autoencoder approach fared in exploring the working space of the robot. Again as a reminder, the autoencoder approach used said autoencoder to reduce the state-action-pair in its dimensionality and thus, with the help of the CMA-ES, allowed to choose the next action, based on the lowest density. Figure ? shows the robot's endeffector position after 60, 120, 160 and 200 episodes for two examples in Cartesian coordinates. In contrast to the baseline approach, we can easily see that robot explored far more area of the working space. Even after 120 episodes, we can already observe that the autoencoder approach has explored the area quite evenly. Furthermore it seems that this method does not keep visiting the same coordinates again and again,

since with increasing episodes the plots become more densely packed (compare the plots after 120 and 200 episodes).

The same can be observed in figure ? as well. It shows the plots of the density of the endeffector's position after 60, 120, 160 and 200 episodes. We can see that in contrast to the baseline approach, the density is not as centered around the mean. It is more evenly distributed between the positive and negative constraint of the joint. Nevertheless, we can also observe that the endeffector (as well as the other joints, see Appendix) seem to often occupy the fringe position, that is the position that is equal to the constraint. It seem that in the sampling process it often is the case that the lowest density can be reached, when one or more joints are locked in place and only the other joints are moving then.

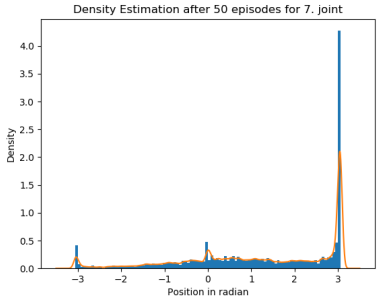
The evolution of the density distribution can be seen in figure ?. It shows the the density of the entire data (state action pairs) after 1, 50, 100 and 150 episodes. Since the goal of this approach is the minimization of exactly this density, the first the plots (after 1 and after 50 episodes) shows what is expected. The density decreased more or less significantly. But when we look at the density after 100 episodes, something unanticipated happened. The density increased. But if we now compare that to the figure ? and figure ?, we can see that already after 100 episodes, the approach has explored the working space quite evenly, which makes it harder for the process to find an action that leads to a position that has not yet been occupied and therefore has low density. Consequently the algorithm will oftentimes choose an action that will lead to an already visited position, what increases the total density.

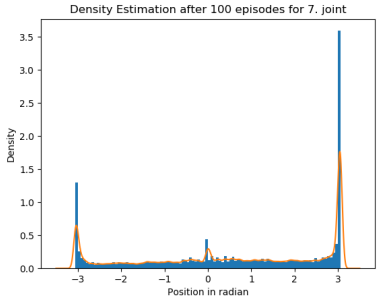


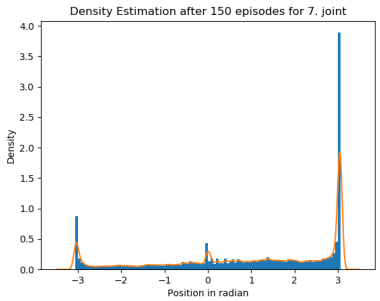
Plot of endeffector of one simulation after 60, 120, 160 and 200 episodes

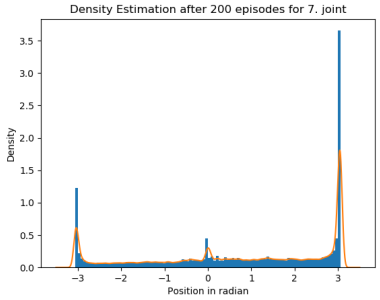
Plot of endeffector of another simulation after 60, 120, 160 and 200 episodes

Figure 11: Comparison of the endeffector position after a different episodes

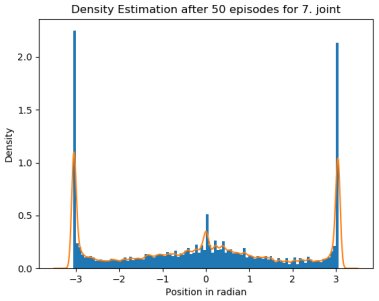


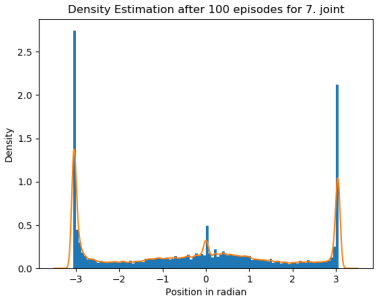


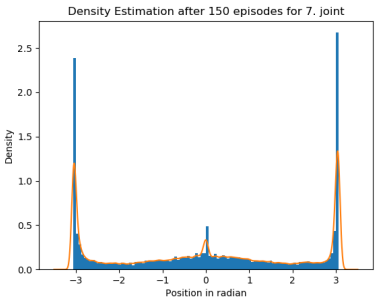


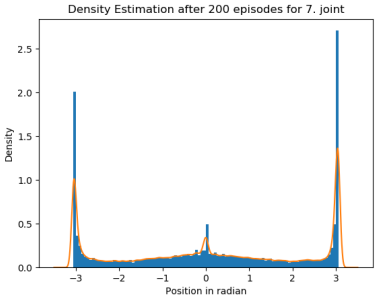


Plot of density of endeffector position of one simulation after 60, 120, 160 and 200 episodes



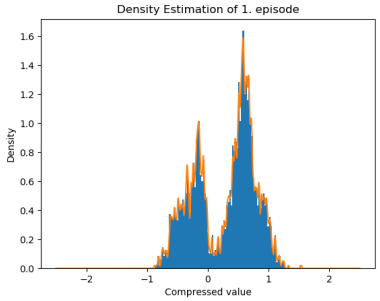


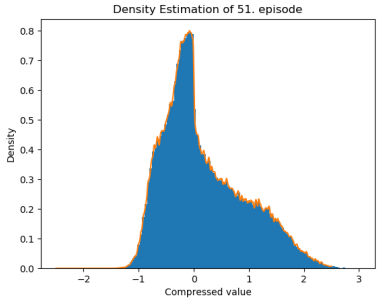


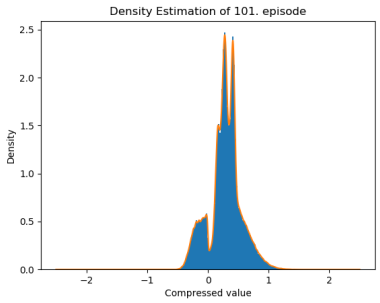


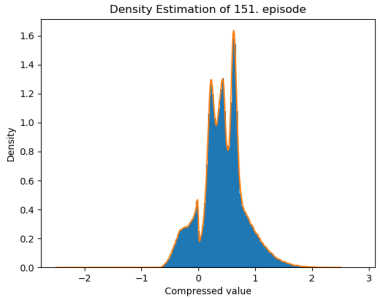
Plot of density of endeffector position of one simulation after 60, 120, 160 and 200 episodes

Figure 12: Comparison of the density of the position of the endeffector

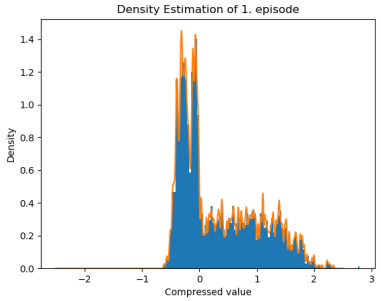


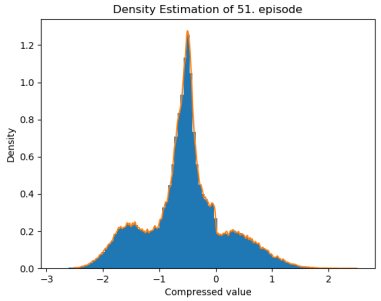


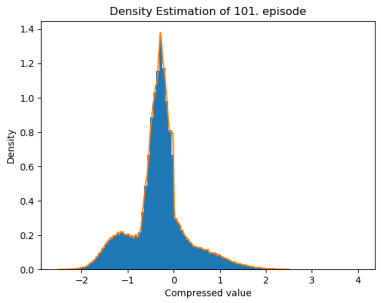


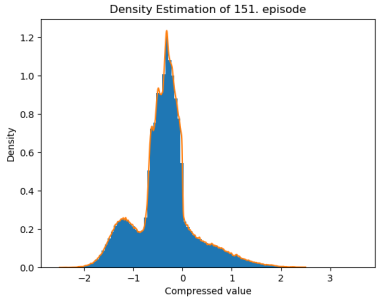


Plot of density of endeffector position of one simulation after 60, 120, 160 and 200 episodes









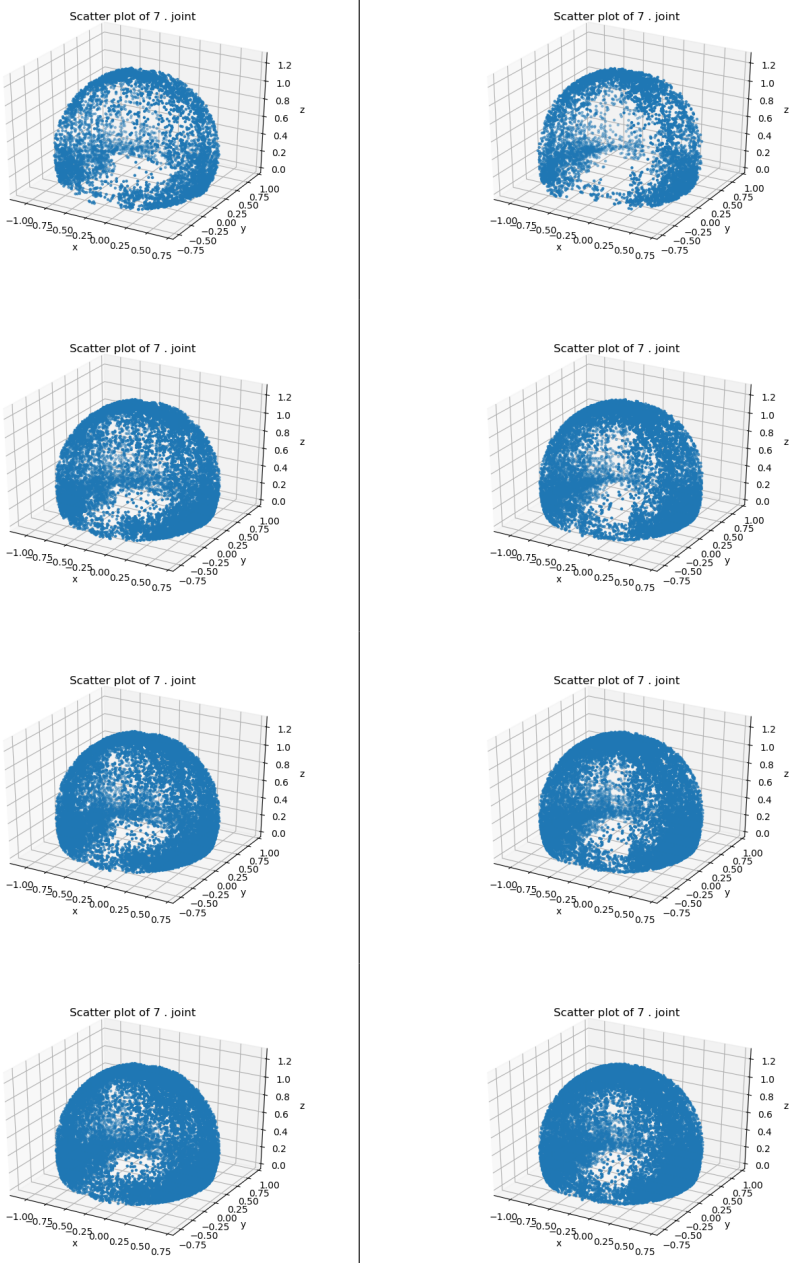
Plot of density of endeffector position of one simulation after 60, 120, 160 and 200 episodes

Figure 13: Comparison of the density of the position of the endeffector

6.3 PCA exploration

Let us now look at the last approach of this thesis. Again, as a reminder, the PCA-approach reduced the dimensionality of the state-action-pair and thus, with the help of the CMA-ES, allowed to choose the next action based on the lowest density. In figure ? we can see the result of that approach. It shows the plots of the position of the robot's endeffector after 60, 120, 160 and 200 episodes for two examples in Cartesian coordinates.In contrast to the baseline approach we can once more see that the PCA approach seems to explore the space much more quickly and evenly. Nevertheless, when we take a closer look and compare it to the results of the autoencoder approach, we can observe that the PCA approach is nowhere close to explore the same area. The entire interior of the half sphere was apparently not visited by the endeffector.

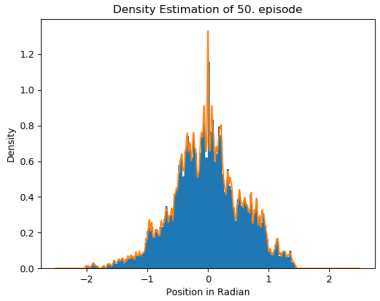
The same conclusion can be drawn from figure ?. It shows the density of the endeffector's inherent position in Radian. Since the density does not spread out evenly between the positive and negative joint constraint and since the same can be observed for all other joints (see Appendix), we can deduce that not the entire working space of the robot has been visited, given that not the entire range of the joints has been used.

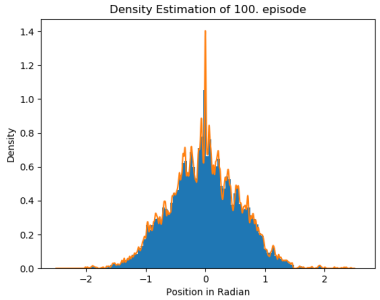


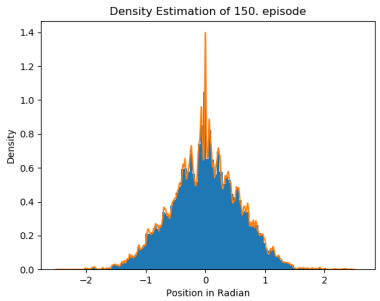
Plot of endeffector of one simulation after 60, 120, 160 and 200 episodes

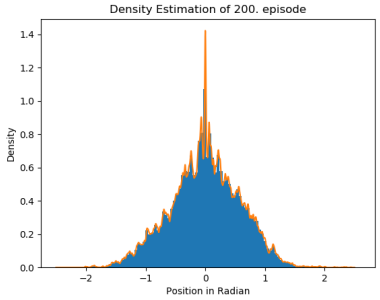
Plot of endeffector of another simulation after 60, 120, 160 and 200 episodes

Figure 14: Comparison of the endeffector position after a different episodes

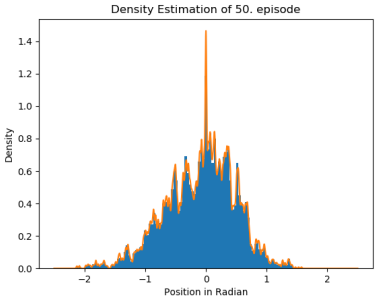


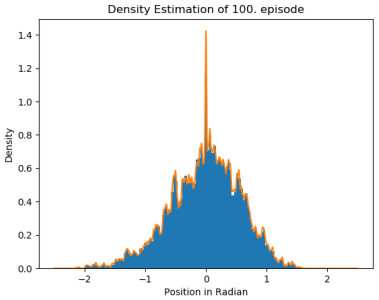


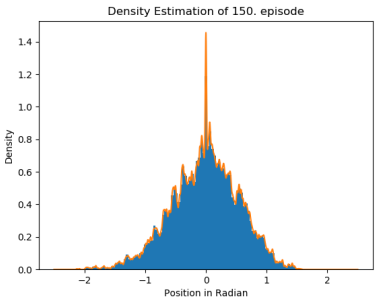


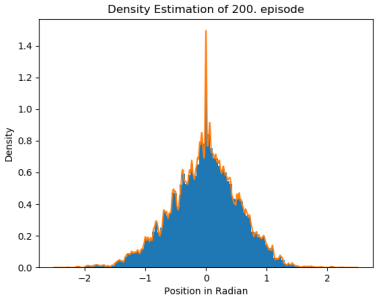


Plot of density of endeffector position of one simulation after 60, 120, 160 and 200 episodes







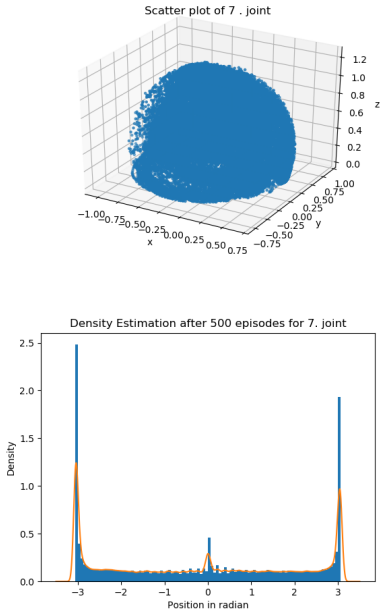


Plot of density of endeffector position of one simulation after 60, 120, 160 and 200 episodes

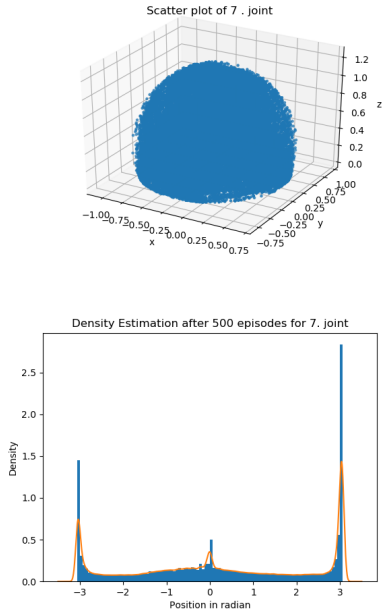
Figure 15: Comparison of the density of the position of the endeffector

6.3.1 Additonal

Each approach has been tested for at least two different simulations with the same settings for 200 episodes. The autoencoder approach however has been conducted for 500 episodes. You can see the additional data in the figure ? below.



Plot of the position of the density of the position of the endeffector after 500 episodes for one simulation



Plot of the position of the density of the position of the endeffector after 500 episodes for another simulation

Figure 16: Comparison of position and the density of the position of the endeffector

7 Conclusion

This thesis presented three approaches for exploring the working space of a robot: Sampling from a normal distribution, sampling with the CMA-ES with density reduction through an autoencoder and sampling with CMA-ES with density reduction through PCA. We can evaluate these approaches through two different methods.
The first would be the time required to complete all 200 episodes. Here the random sampling from a normal distribution took the least and a constant amount of time, because the calculation was the same in each iteration and

the computational cost did not rise with an increase in episodes. So it took roughly 200*100*1 seconds for the program to finish. The PCA approach was the second fastest. The most time intensive computation here was the fitting of the PCA with the data. To limit the computational cost, this data was restricted to the last 20 episodes. So, after these 20 episodes, each iteration took about 1.5 seconds. Hence the entire time is roughly 200*100*1.5 seconds. The autoencoder approach the most amount of time. In each iteration the existing autoencoder had to be evaluated on the collected data. Like with the previous approach, to limit computational time, this data has been restricted to the last 20 episodes. After these an iteration took about 1.5 seconds. Thus, the computational time during the episode is similar to PCA approach. But after each episode it was also necessary to train the autoencoder. That was done with all the to this point acquired data. Hence the computational time increased with the increase of the episodes. While in the first few episodes the training did not take make than a few seconds, in the later episodes the same training took almost an hour.
The first method for evaluation is how well the space has been explored. Each approach ran for 200 episodes, which means that the amount of occupied points is the same. Now we can look how evenly they are distributed though out the space. For that we can look onto the figures from the previous chapter, namely figure ?, ? and ?. Here we can see that the random sampling from a normal distribution fared the worst. Only a chunk of the entire space has been explored and it seems that the robot's endeffector keeps visiting the same spots over and over. The other two approaches seem to explore the space much more evenly. But we can clearly see, when we compare figure ? and figure ? that the autoencoder approach was exploring the working space of the robot more extensively than the PCA approach, for the interior of the sphere has not been well explored in the latter.