The main concern of all reviewers was that the paper was hard to follow in general. This was addressed by re-writing the Introduction, Background and Methods section such to be more compressive.

To give a clear overview of the entire learning and data flow Figure 4 has been added. I shows side by side the information flow with all the variables defined whilst also detailing the learning procedure. Each of the elements in the figure “Data collection”, “Learning” and “Control” have their own sections in the text and are referenced throughout for clarity.

<u>Response to reviewer 1</u>

The suggestion of the reviewer was to re-write the paper such to make the overall structure easier to following. The addition of Figure 4 and the rewriting of the Methods section addresses this issue.

<u>minor remarks</u>

1) authors use the same notation (x and y) for position vector and binary feature vector. The same notation was latter used to denote Cartesian coordinate axis.

Addressed.

2) Figure 2: Likelihood edge contact is not clear to me. Additional explanation would be beneficial.

3) page 4: Authors referee to the Figure 4 bottom, while only Figure 4 top exists.

Addressed.

4) page 6: “Maximization EM step, see Figure 2, is obtained.” Should be Figure 5 instead??

Addressed, now point to the correct figure.

5) page 7: Figure 7: The scale of the value function in the range of 0..100, while in the text is in the range of 0..1.

The value function with scale 0-1 is associated with the 2D MDP example, whilst the value function with scale 0-100 is associated with POMDP PiH problem. This has been revised to be clear in the text.

6) page 8: Eq. (10) is ambiguous, has the same value (\dot{x}) on both sides. Please rewrite it.

Has been addressed (note that now \dot{x} = u)

7) Experiments considered only position velocity learning. I suggest including also rotational velocity learning in experiments or give an explanation, why this was omitted.

The uncertainty in the PiH experiment was only present in the transitional location of the socket and not its orientation. The main novelty of the paper was the introduction of the Fitted Policy Iteration algorithm and its strengths were easier to highlight for the linear velocity policy. The angular velocity only comes into play right at the end of the task. As a result doing a detailed analysis on the angular velocity component would not bring much to the paper as it has already been done in other publications.

Response to reviewer 2

major

One of my main concerns about this paper is the way it is written. Unfortunately I have to say that the reading flow is not smooth and therefore it is quite hard to follow the details and idea of the paper. The writing is lousy and several mistakes were found (some of them are detailed at the end of this review). This, of course, significantly reduces the quality of this work and makes the material presented in the paper less accessible to a broader audience. So, my first suggestion is to carefully review the writing/notations/concepts of the paper so that the text fulfills the high-standard requirements of the journal to which this paper is submitted to.

The introduction, background have had major re-write. The notation has been changed to be consistent throughout. To ease following of the concepts and what is done Figure 4 was added.

- Section Experiment methods:

What is the motivation of using differentiation entropy to compress the belief PDF? What is the intuition behind this? For example, why not using latent representations if the aim is reducing the dimensionality of the space in which the search is carried out?

This is a possibility, however it would require a discretization of the state space and would be slow. For instance every time step the PMF would have to be fitted to the discrete representation before being transformed to a lower dimensional representation. I have experimented with E-PCA vs MLS and Entropy in a 2D POMDP navigation task. I found that if the behaviour demonstrated by teachers is consistent then there is little difference in the effect the compression method has. E-PCA is however more advantageous if random exploration strategies are used during the learning of the policy.

- Section Learning Actor & Critic

+ The authors say that "Both the linear and angular velocity policies are parameterised by a Gaussian Mixture Model (GMM)" However, the way how the GMM parameters are defined later describes means and covariance matrices encoding linear velocities and belief space vectors 'b'. This is really confusing because it is totally unclear if angular velocities are included in \dot{x} (according to the notation this should not be the case), or if the authors made a mistake here and they are encoding only linear velocities and belief space vectors 'b'.

Addressed, the notation of the entire document has been updated to avoid any conflict.

+ I found that Algorithm 1 does not really add something new or different to the description of the Fitter Policy Iteration method, neither facilitates the understanding of this technique. However, I encourage the authors to use this Algorithm to show how the whole learning process is carried out (1. Demonstrations are collected. 2. Belief PDF is approximated by PMF. 2.1. Differentiation entropy is used to compress the belied PDF, etc...)

I have left the algorithm as it is, but I have added Figure 4 which should address these issues.

+ Please explicitly indicates that Q is a function that represents the logarithmic lower bound. This is not indicated in the paper.

Addressed.

- Section Results:

+ Good set of experiments. However, it may be useful to mention how other policy search algorithms may perform here. My point is that greedy search is used as a sort of baseline for comparison purposes. However, as part of the contributions of this paper is about Q-EM, it would be more interesting to see how other policy search methods behave in the same scenarios.

In terms of other reinforcement learning methods I do not know of which could address this problem efficiently. Policy Search methods work well for reactive or motion primitive like behavior. The reason being that these policies do not have many parameters which would not be the case for the PiH localisation and connection task. Also most of the RL examples applied to robotics consider time dependent dynamical systems (such as DMP) which would not be ideal for a long search process. A method like PoWER could also be considered but requires many rollouts for each new set of parameters. But I agree, it would have been nice to compare with another approaches which is not as naive as a myopic policy.

+ I wonder if it is really fair to compare an algorithm that does not include any information about the goal of the task (GMM) while the other has additional information to work with (Q-EM)? In other words, the training algorithm of Q-EM does include information about the task goal, which is not given to the GMM (I might miss something here if this is not the case), therefore, in my opinion, as soon as Q-EM converges, it will provide better results that GMM (even better if the demonstrations are not describing the task goal in a good way).

The GMM algorithm implicitly knowns location of the goal because statistically all the episodes will end at the same location (socket-plug connection). As a result the GMM will reproduce the same behavior. The Q-EM algorithm only has one bit of more information than the GMM which is the binary reward function. I only did one iteration of policy evaluation and improvement and did not do any additional rollouts to keep the comparison between GMM and Q-EM fair. It is true that if multiple iterations of the Q-EM algorithm was performed you would be correct.

- Section Discussion & Conclusion:

This looks more like a summary of the paper than a discussion. Nothing important or relevant is added here!

I agree, it has been rewritten to emphasis the key contribution and what future work could be done.

Response to reviewer 3

The main concern of the reviewer lies in the structure of the paper and the fields chosen to be covered in the background.

Following the suggestion of the reviewer the background was changed to address more specifically the POMDP literature.

Section 3.2 and was moved before 3.1 and Section 3 and 4 have been unified.