# Imitation and Reinforcement Learning from Failed Demonstrations

**Daniel H Grollman** DANIEL.GROLLMAN@EPFL.CH
**Aude G Billard** AUDE.BILLARD@EPFL.CH
Learning Algorithms and Systems Laboratory
Ecole Polytechnique Fédérale de Lausanne

## Abstract

Current work in robotic imitation learning uses successful demonstrations of a task performed by a human teacher to initialize a robot controller. Given a reward function, this learned controller can then be improved using techniques derived from reinforcement learning. We instead use failed attempts, which may be more plentiful, to initialize our controller and, taking them as illustrations of what *not* to do, deliberately generate behaviors that differ from the human's.

## 1. Introduction

Given a physically capable robot, **Robot Learning from Demonstration** (RLfD) seeks to generate autonomous controllers from observation instead of programming (Argall et al., 2009). Current state of the art techniques first use one or more successful executions of a task performed by a human teacher to initialize the controller and then use an explicit reward function in a self-improvement phase to optimize the controller via robot practice.

Instead, we seek to initialize our learning by observing failed attempts on the part of the human. In doing so we are inspired by Meltzoff (1995), who demonstrated that human babies are capable of learning tasks from only observing failed executions. Part of our inspiration also comes from the fact that in demonstrating a task to a robot, the human may generate several failed trials before providing one suitable for success-based RLfD. This data is usually discarded, and the amount of time spent acquiring it is unreported. However, by utilizing this data we may be able to improve learning efficacy when we consider not only the time spent by the robot learning (which may be greater when learning from failure), but also the time spent demonstrating by the human (which may be less when failed demonstrations are allowed).

In learning from failure, we make two assumptions:

1. The human is attempting to perform the task.
2. The human attempts to correct for failures.

These two assumptions lead us to an approach to the credit assignment problem, by positing that the variance between failed demonstrations is indicative of how correct different portions of the demonstration are. If all demonstrations agree on a part of the behavior (low variance), we take it that it is likely to be performed correctly. On the other hand, if all demonstrations disagree on a part (high variance), we assume that it is likely to be incorrectly performed in all of them, and the multiple errors indicate the human's intuition as to what values should be explored.

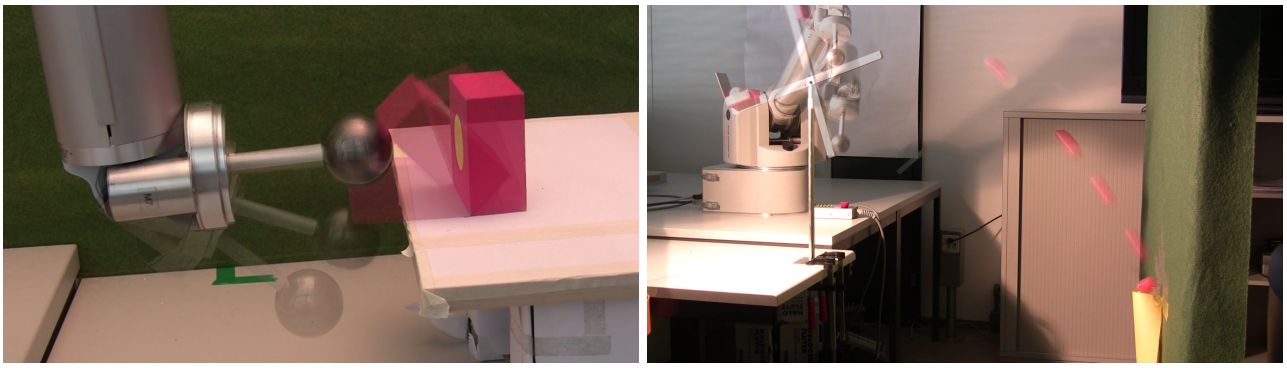
## 2. Models

We model motions as Autonomous Dynamical Systems (ADS) (Hersch et al., 2008) where velocity is computed as a function of the current position: $\dot{\xi} = f(\xi|\theta)$. The parameters to this function $\theta = \{K, \{\rho^k, \mu^k, \Sigma^k\}_{k=1}^K\}$ define the Gaussian Mixture Model (GMM) distribution over the joint:

$$P(\dot{\xi}, \xi|\theta) = \sum_{k=1}^{K} \rho^k \mathcal{N}(\dot{\xi}, \xi|\mu^k, \Sigma^k) \qquad (1)$$

where the priors ($\rho^k$) sum to one and $\mathcal{N}$ is the normal distribution with mean $\mu^k$ and variance $\Sigma^k$. These parameters are fit to demonstrated state-velocity pairs using weighted versions of the Expectation Maximization (Neal & Hinton, 1998) and Bayesian Information Criterion techniques (Hu & Xu, 2004).

Appearing in *Proceedings of the ICML Workshop on New Developments in Imitation Learning*, Bellevue, WA, USA, 2011. 



(a) Flip Up (b) Basket

*Figure 1.* The FlipUp Task (left): Get the foam block to stand on end. The Basket task (right): Launch the ball into the basket. Shown are successful performances learnt from failure

New trajectories are generated from an initial state $\xi_0$ by computing the velocity $\dot{\xi}_0 = f(\xi_0|\theta)$, moving with that velocity, and repeating to quiescence or timeout.

### 2.1. PoWER

The PoWER (Policy learning by Weighting Exploration with the Return) algorithm of Kober & Peters (2010) generates exploratory $\theta'$ given the $N$ most highly rewarded previously tried parameters and their rewards $\{\phi^n, R^n\}_{n=1}^N$. The mean parameters maximize a lower bound on the expected reward:

$$\bar{\theta} = \frac{1}{\sum_{n=1}^N R^n} \sum_{n=1}^N R^n \phi^n \tag{2}$$

and are corrupted by mean-zero Gaussian noise independently on each parameter, $\theta' \sim \mathcal{N}(\bar{\theta}, \sigma^2 \mathbf{I})$.

Generated velocities are the expectation of the conditional distribution, $\dot{\xi} = f(\xi|\theta) = E\{P(\dot{\xi}|\xi, \theta)\})$.

### 2.2. Donut

Grollman & Billard (2011) present an alternative method of generating motions from a GMM, specifically designed for the failed demonstration case. It replaces each of the conditional Gaussian distributions $\mathcal{N}(\dot{\xi}|\xi, \mu^k, \Sigma^k)$ with a so-called Donut distribution, $\mathcal{D}(\dot{\xi}|\xi, \mu^k, \Sigma^k, \epsilon)$. This distribution is a center-off distribution whose width is controlled by the additional exploration parameter ($\epsilon$) up to a maximum $\lambda^*$ as shown in Figure 2. The velocity for a given state is:

$$\dot{\xi} = f(\dot{\xi}|\xi, \theta) = \text{argmax}_{\dot{\xi}} \sum_{k=1}^K \rho^k \mathcal{D}(\dot{\xi}|\xi, \mu^k, \Sigma^k, \epsilon) \tag{3}$$

and found by gradient ascent (starting at the current $\dot{\xi}$). Exploration is set as $\epsilon = 1 - (1 + ||V\{\dot{\xi}|\xi, \theta\}||)^{-1}$. By this method, areas that are consistent across demonstrations are replicated, while those that are varied by the demonstrator are explored more.

We extend this method here to incorporate reward information by weighing each datapoint in a trajectory $(\xi_t^s, \dot{\xi}_t^s)$ by the reward for the entire trajectory, $R^s$ (associated with parameters $\theta^s$). We then re-estimate parameters $\theta^{S+1}$ from all of the reward-weighted datapoints using weighted EM. We do not re-estimate $K$ for a fairer comparison with PoWER.

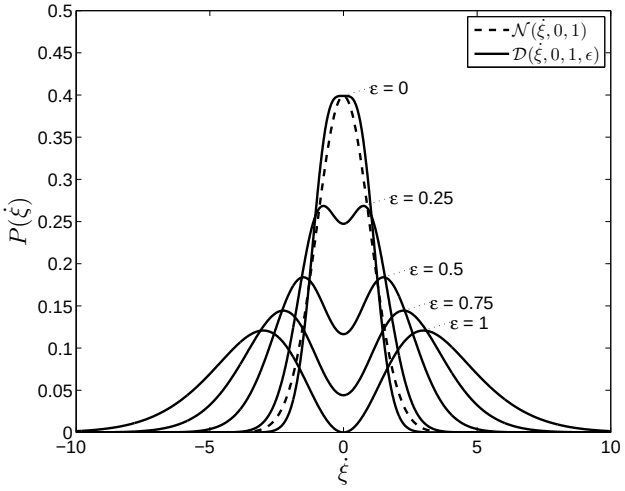



*Figure 2.* The Donut distribution for various values of $\epsilon$. Also shown is the Normal distribution for comparison.

## 3. Experiments

We compare PoWER and Donut on two robot tasks, shown in Figure 1. In the FlipUp task (Figure 1(a)) we consider two failure cases: If the block does not pass the target position and falls back, reward is measured as $R = \exp(-\text{argmin}_{\gamma^t}|\gamma^t|)$, with $\gamma$ being the angle of the block with respect to the surface normal of the table. If the block instead passes to the other side, reward is measured as $R = \exp(-6|\dot{\gamma}^{t^*}|)$, where $t^*$ is the time at which the block passes the upright position.

For the Basket case (Figure 1(b)), reward is computed as $R = \exp(-|y|)$, where $y$ is the vertical offset of the ball from the lip of the basket when it makes contact with the wall. For both rewards the necessary information is extracted from a fast stereo vision pair.

In our experiments, we initialize the GMM for each task with two failed kinesthetic demonstrations. The robot then generates trials autonomously, and a human observer stops it when success is achieved. Our results are summarized in Table 1, showing the means and standard deviations of the number of trials to achieve success over multiple initial training sets (10 for FlipUp, 3 for Basket). We also provide averages over multiple humans for comparison.

## 4. Discussion and Future Work

Analyzing our results, we see only minor differences in the average number of trials to succeed in both algorithms. Where we do find a difference, however, is in the variance, that of Donut is nearly an order of magnitude smaller than that of PoWER. We believe this result is due to the targeted method by which Donut explores. While PoWER varies all portions of the behavior equally, Donut instead focuses on the areas of unsurity in the demonstration. A similar behavior in PoWER may be possible using alternate exploration parameters, perhaps by initializing them based on the demonstration variance.

Looking forward, we note that currently, PoWER operates on the parameters of a motion, while Donut works directly on generated velocities. However, PoWER itself can be viewed as building a GMM in the parameter space, so we can consider *lifting* Donut to work in the parameter space instead. Doing so should allow us to work in higher dimensional spaces, and also address two issues with the current approach: the time consuming gradient ascent step, and the non-guarantee of smooth motions.

| | FlipUp | Basket |
|---|---|---|
| **Donut** | $4.30 \pm 0.48$ | $7.67 \pm 0.58$ |
| **PoWER** | $4.60 \pm 2.17$ | $11.00 \pm 5.29$ |
| **Human** | $5.2 \pm 3.11$ | $3.50 \pm 1.73$ |

*Table 1.* Summary of results

## Acknowledgements

This work has partially been supported by the European Commission under contract numbers FP7-248258 (First-MM) and FP7-ICT-248311 (Amarsi). Florent D'Halluin was of great assistance in carrying out the experiments described herein.

## References

Argall, Brenna D., Chernova, Sonia, Veloso, Manuela, and Browning, Brett. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469 – 483, May 2009.

Grollman, Daniel H and Billard, Aude. Donut as I do: Learning from failed demonstrations. In *International Conference on Robotics and Automation*, Shanghai, May 2011.

Hersch, Micha, Guenter, Florent, Calinon, Sylvain, and Billard, Aude. Dynamical system modulation for robot learning via kineshetic demonstrations. *IEEE Transactions on Robotics*, pp. 1463–1467, 2008.

Hu, Xuelei and Xu, Lei. Investigation on several model selection criteria for determining the number of cluster. *Neural Information Processing. - Letters and Reviews*, 4(1):1–10, July 2004.

Kober, Jens and Peters, Jan. Policy search for motor primitives in robotics. *Machine Learning*, 2010.

Meltzoff, Andrew N. Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31(5):838–850, 1995.

Neal, Radford and Hinton, Geoffrey E. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pp. 355–368. Kluwer Academic Publishers, 1998.