Robot Learning

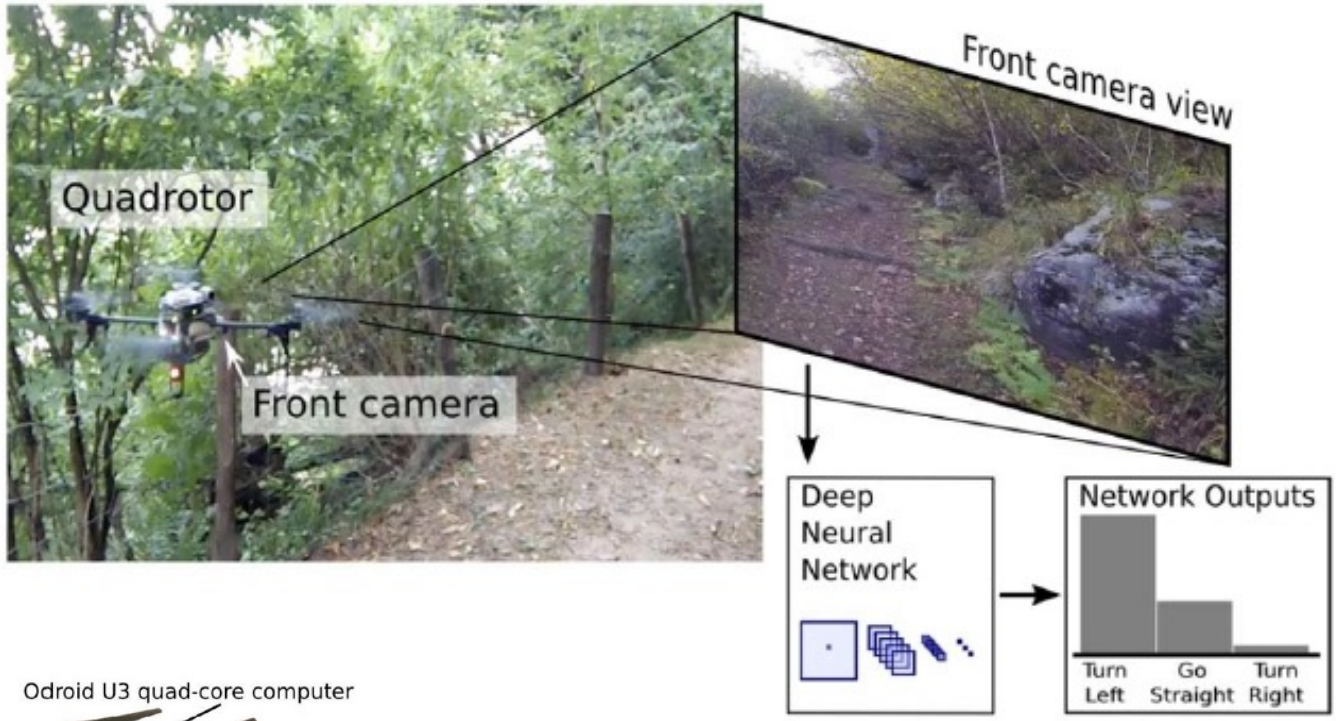
Behavior Cloning, DAgger

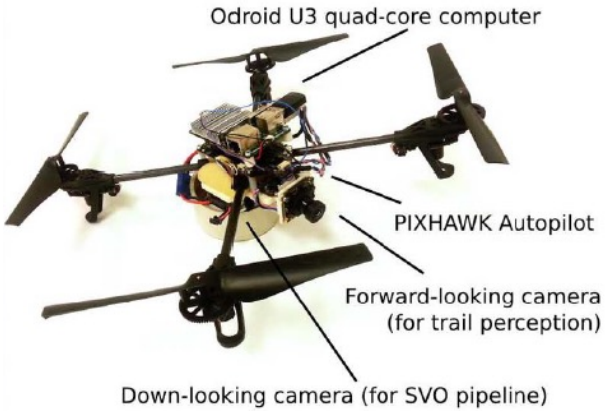
Prof. Songhwai Oh
ECE, SNU

IEEE Robotics and Automation Letters, vol. 1, no. 2, pp. 661-667, July 2016

A Machine Learning Approach to Visual Perception of Forest Trails for Mobile Robots

Approach

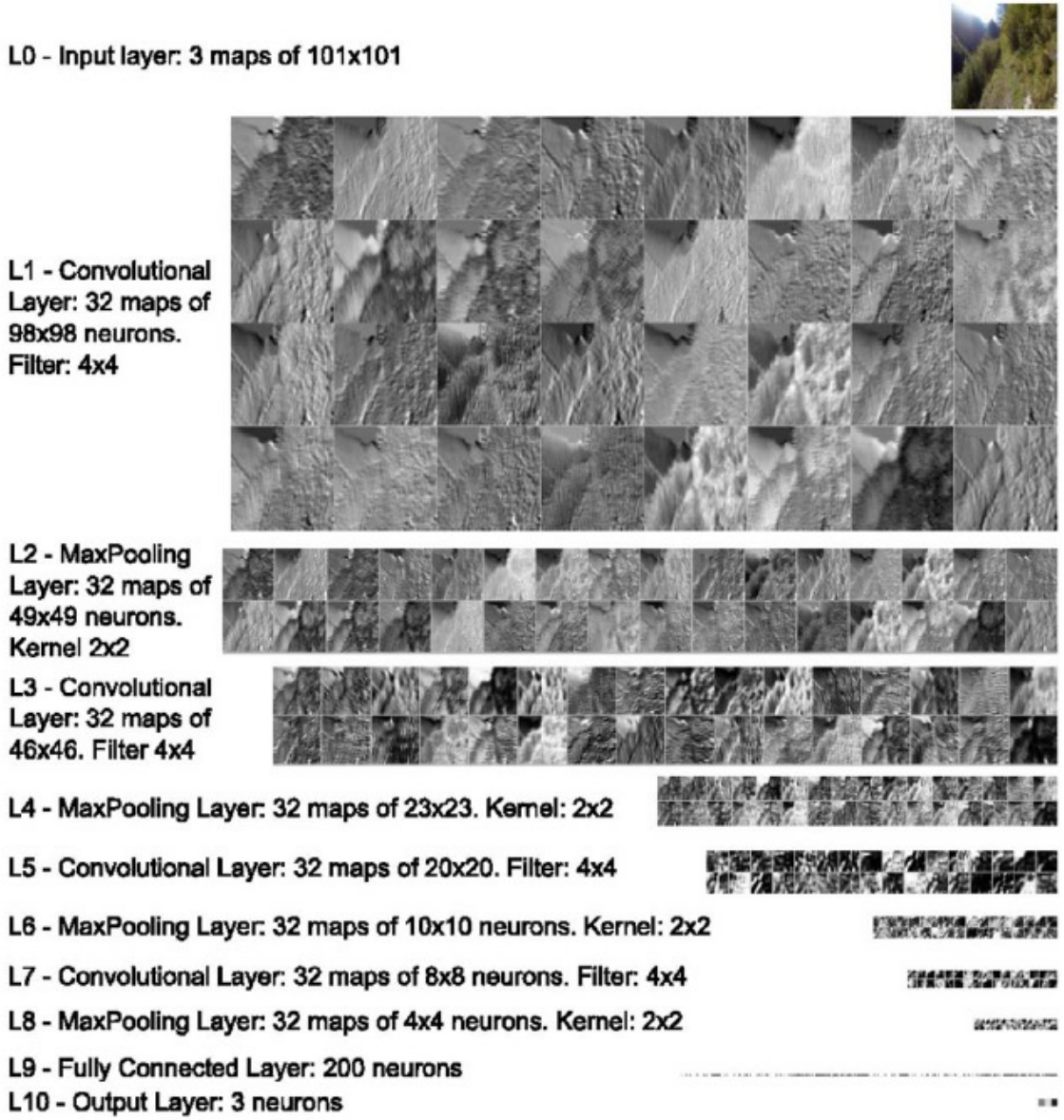




Classification problem:

- Input: Image
- Output: Turn Left, Go Straight, Turn Right

CNN Architecture



Data Collection

Fig. 4. *Left:* stylized top view of the acquisition setup; *Right:* our hiker during an acquisition, equipped with the three head-mounted cameras.

Video

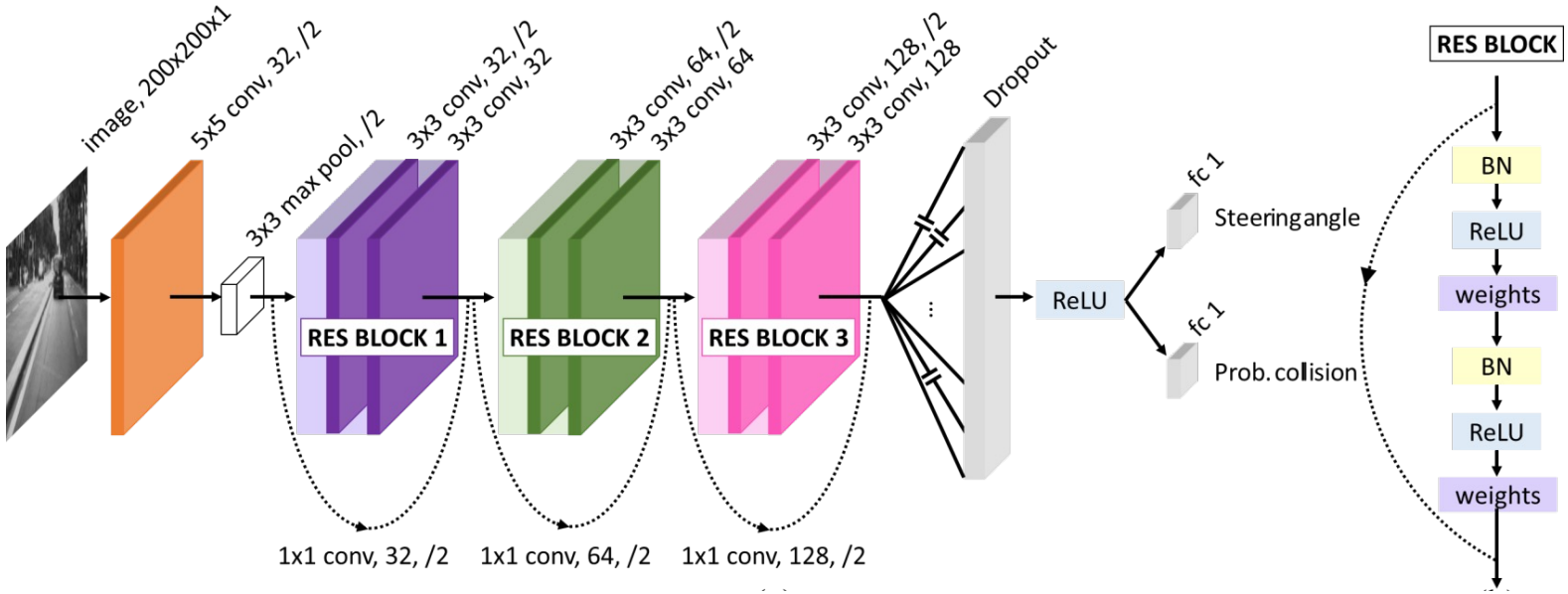
IEEE Robotics and Automation Letters, vol. 3, no. 2, pp. 1088-1095, April 2018

DroNet: Learning to Fly by Driving

Approach

Driving Data (Udacity)

Bicycle Data (Collected)



Loss: $L_{tot} = L_{MSE} + \max(0, 1 - \exp^{-decay(epoch - epoch_0)}) L_{BCE}$

Udacity Data

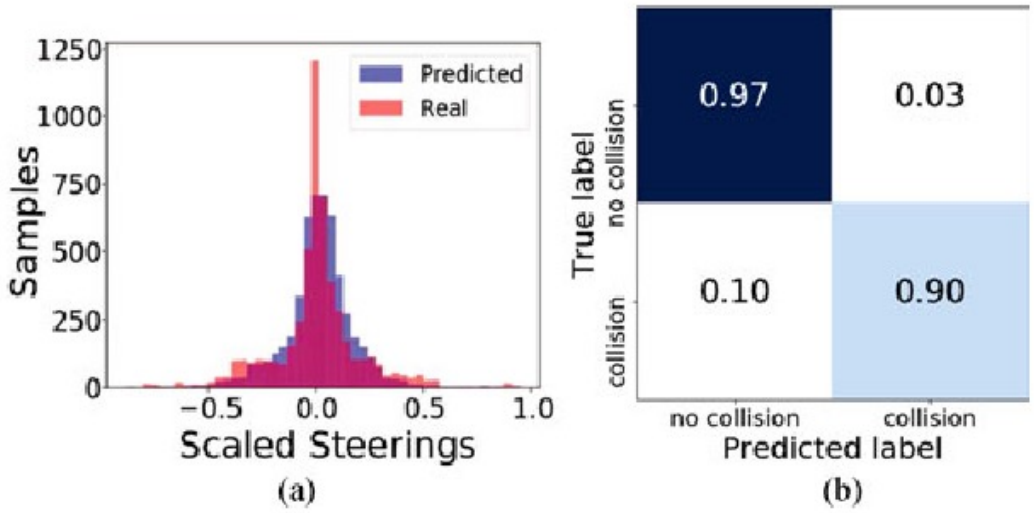
Bicycle Data

Results

TABLE I
QUANTITATIVE RESULTS ON REGRESSION AND CLASSIFICATION TASK: EVA AND RMSE ARE COMPUTED ON THE STEERING REGRESSION TASK, WHILE AVG. ACCURACY AND F-1 SCORE ARE EVALUATED ON THE COLLISION PREDICTION TASK

Model	EVA	RMSE	Avg. accuracy	F-1 score	Num. Layers	Num. parameters	Processing time [fps]
Random baseline	-1.0 ± 0.022	0.3 ± 0.001	$50.0 \pm 0.1\%$	0.3 ± 0.01	–	–	–
Constant baseline	0	0.2129	75.6%	0.00	–	–	–
Giusti *et al.* [9]	0.672	0.125	91.2%	0.823	6	5.8×10^4	23
ResNet-50 [18]	0.795	0.097	96.6%	0.921	50	2.6×10^7	7
VGG-16 [23]	0.712	0.119	92.7%	0.847	16	7.5×10^6	12
DroNet (Ours)	0.737	0.109	95.4%	0.901	8	3.2×10^5	20

EVA: Explained variance ratio
RMSE: Root mean squared error



Video


DroNet: Learning to Fly by Driving
Antonio Loquercio, Ana I. Maqueda, Carlos R. del Blanco and Davide Scaramuzza
University of Zurich[UZH]
Department of Neuroinformatics
ETH zürich
University of Zurich[UZH]
Department of Informatics
POLITÉCNICA

International Conference on Artificial Intelligence and Statistics, 2011

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning

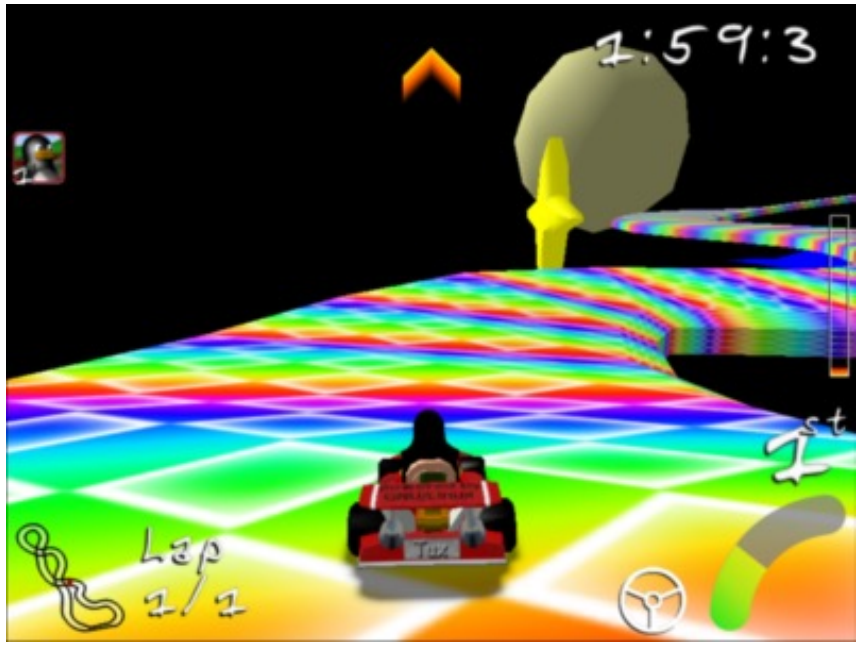
DAgger (Dataset Aggregation)

Initialize $\mathcal{D} \leftarrow \emptyset$.
Initialize $\hat{\pi}_1$ to any policy in Π.
for $i = 1$ **to** N **do**
  Let $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$.
  Sample T-step trajectories using π_i.
  Get dataset $\mathcal{D}_i = \{(s, \pi^*(s))\}$ of visited states by π_i and actions given by expert.
  Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$.
  Train classifier $\hat{\pi}_{i+1}$ on $\mathcal{D}$.
end for
Return best $\hat{\pi}_i$ on validation.

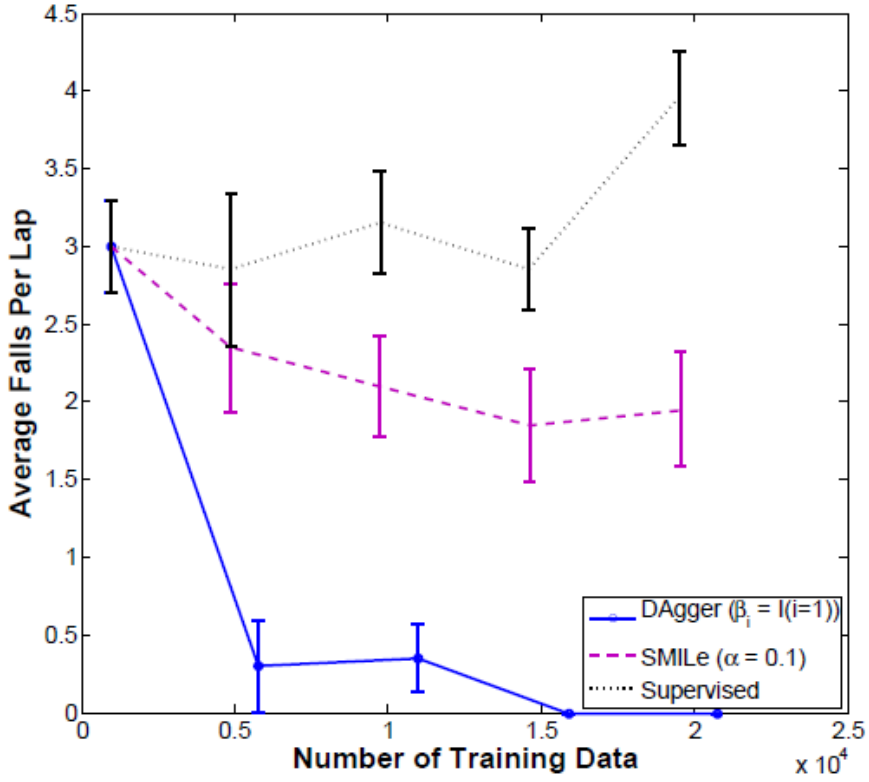
π^*: expert's policy
$\beta_1 = 1, \beta_i \leq (1 - \alpha)^{i-1}$

Intuition: Building up the set of inputs that the learned policy is likely to encounter during its execution based on previous experience (training iterations).

Experiment: Super Tux Kart

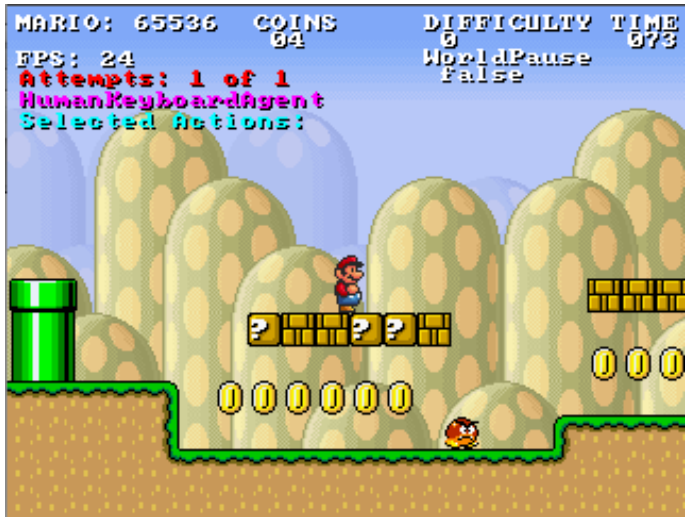


[4]Features x: LAB color values of each pixel in a 25x19 resized image of the 800x600 image; output steering: $\hat{y} = w^T x + b$ where w, b minimizes ridge regression objective: $L(w,b) = \frac{1}{n}\sum_{i=1}^{n}(w^T x_i + b - y_i)^2 + \frac{\lambda}{2} w^T w$, for regularizer $\lambda = 10^{-3}$.

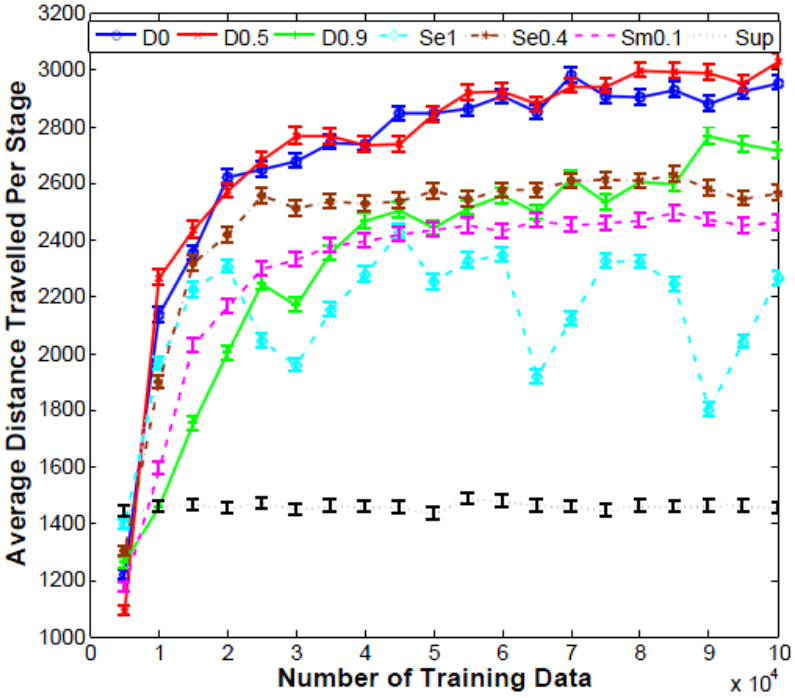


Supervised: All training examples are similar and do not help the learner to learn how to recover from mistakes it makes

Experiment: Super Mario Bros.



[5]For the input features x: each image is discretized in a grid of 22x22 cells centered around Mario; 14 binary features describe each cell (types of ground, enemies, blocks and other special items); a history of those features over the last 4 images is used, in addition to other features describing the last 6 actions and the state of Mario (small,big,fire,touches ground), for a total of 27152 binary features (very sparse). The k^{th} output binary variable $\hat{y}_k = I(w_k^T x + b_k > 0)$, where w_k, b_k optimizes the SVM objective with regularizer $\lambda = 10^{-4}$ using stochastic gradient descent (Ratliff et al., 2007; Bottou, 2009).



Problem Setup

- Π: set of policies, where $\pi \in \Pi$ is a policy
- T: task horizon
- d_π^t: distribution of states at time t with policy π executed from time 1 to time $t-1$
- $d_\pi = \frac{1}{T}\sum_{t=1}^{T} d_\pi^t$ (state visitation frequency)
- $C(s,a)$: immediate cost $(C(s,a) \in [0,1])$
- $C_\pi(s) = \mathbb{E}_{a\sim\pi(s)}\left(C(s,a)\right)$
- $J(\pi) = \sum_{t=1}^{T} \mathbb{E}_{s\sim d_\pi^t}\left(C_\pi(s)\right)$ (cost-to-go)
- π^*: expert's policy
- $l(s,\pi)$: loss function e.g., $l(s,\pi) = \mathbf{1}\left(\pi(s) \neq \pi^*(s)\right)$
- $C_\pi(s) = l(s,\pi)$ in this paper

Goal: $\hat{\pi} = \arg\min_{\pi\in\Pi} \mathbb{E}_{s\sim d_\pi}\left(l(s,\pi)\right)$

Supervised Learning

Supervised learning: $\pi_{sup} = \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_{\pi^*}} (l(s, \pi))$

Theorem 2.1. (Ross and Bagnell, 2010) Let $\mathbb{E}_{s \sim d_{\pi^*}} (l(s, \pi)) = \epsilon$, then $J(\pi) \leq J(\pi^*) + T^2 \epsilon$.

(Proof Sketch)

- Let $\epsilon_i = \mathbb{E}_{s \sim d^i_{\pi^*}} (l(s, \pi))$, the expected loss at time i of π. Then $\epsilon = \frac{1}{T} \sum_{i=1}^{T} \epsilon_i$.
- At time t, the probability of making an error $(\pi(s) \neq \pi^*(s))$ is bounded by $\sum_{i=1}^{t} \epsilon_i$.
- For length T, $\sum_{t=1}^{T} \sum_{i=1}^{t} \epsilon_i \leq T \sum_{t=1}^{T} \epsilon_t = T^2 \epsilon$

Note: The bound is tight.

Theorem 2.2. *Let π be such that $\mathbb{E}_{s\sim d_\pi}[\ell(s,\pi)] = \epsilon$, and $Q^{\pi^*}_{T-t+1}(s,a) - Q^{\pi^*}_{T-t+1}(s,\pi^*) \leq u$ for all action a, $t \in \{1,2,\ldots,T\}$, $d^t_\pi(s) > 0$, then $J(\pi) \leq J(\pi^*) + uT\epsilon$.*

Proof. We here follow a similar proof to Ross and Bagnell (2010). Given our policy π, consider the policy $\pi_{1:t}$, which executes π in the first t-steps and then execute the expert π^*. Then

$$
\begin{aligned}
& J(\pi) \\
&= J(\pi^*) + \textstyle\sum_{t=0}^{T-1}[J(\pi_{1:T-t}) - J(\pi_{1:T-t-1})] \\
&= J(\pi^*) + \textstyle\sum_{t=1}^{T} \mathbb{E}_{s\sim d^t_\pi}[Q^{\pi^*}_{T-t+1}(s,\pi) - Q^{\pi^*}_{T-t+1}(s,\pi^*)] \\
&\leq J(\pi^*) + u\textstyle\sum_{t=1}^{T} \mathbb{E}_{s\sim d^t_\pi}[\ell(s,\pi)] \\
&= J(\pi^*) + uT\epsilon
\end{aligned}
$$

The inequality follows from the fact that $\ell(s,\pi)$ upper bounds the 0-1 loss, and hence the probability π and π^* pick different actions in s; when they pick different actions, the increase in cost-to-go $\leq u$. □

$$\sum_{t=0}^{T-1} \left(J(\pi_{1:T-t}) - J(\pi_{1:T-t-1})\right) = J(\pi_{1:T}) - J(\pi_{1:T-1})$$

$$+ J(\pi_{1:T-1}) - J(\pi_{1:T-2})$$

$$\vdots$$

$$+ J(\pi_{1:1}) - J(\pi_{1:0})$$

$$= J(\pi_{1:T}) - J(\pi_{1:0})$$

$$= J(\pi) - J(\pi^*)$$

telescoping sum

$$J(\pi) = J(\pi^*) + \sum_{t=0}^{T-1} \left(J(\pi_{1:T-t}) - J(\pi_{1:T-t-1})\right)$$

$$\sum_{t=0}^{T-1} \left(J(\pi_{1:T-t}) - J(\pi_{1:T-t-1})\right) = \sum_{t=1}^{T} \left(J(\pi_{1:T-t+1}) - J(\pi_{1:T-t})\right) =: A$$

$Q_t^{\pi^*}(s,\pi)$: t-step cost with initial state s using π then following π^*

$$t = T \qquad J(\pi_{1:1}) - J(\pi_{1:0}) = \mathbb{E}_{s\sim d_\pi^1}\left(Q_T^{\pi^*}(s,\pi) - Q_T^{\pi^*}(s,\pi^*)\right)$$

$$t = T-1 \quad J(\pi_{1:2}) - J(\pi_{1:1}) = \mathbb{E}_{s\sim d_\pi^2}\left(Q_{T-1}^{\pi^*}(s,\pi) - Q_{T-1}^{\pi^*}(s,\pi^*)\right)$$

$$\cdots$$

$$A = \sum_{t=1}^{T} \mathbb{E}_{s\sim d_\pi^t}\left(Q_{T-t+1}^{\pi^*}(s,\pi) - Q_{T-t+1}^{\pi^*}(s,\pi^*)\right)$$

$$\leq \sum_{t=1}^{T} \mathbb{E}_{s\sim d_\pi^t}\left(l(s,\pi)\right) u$$

No-Regret Algorithm

No-regret algorithm generates a sequence of policies $\pi_1, \pi_2, \ldots, \pi_N$, such that

$$\frac{1}{N}\sum_{i=1}^{N}\ell_i(\pi_i) - \underbrace{\min_{\pi\in\Pi}\frac{1}{N}\sum_{i=1}^{N}\ell_i(\pi)}_{\text{Best policy's loss}} \leq \gamma_N \qquad \text{for } \lim_{N\to\infty}\gamma_N = 0.$$

$$\ell_i(\pi) = \mathbb{E}_{s\sim d_{\pi_i}}[\ell(s,\pi)]$$

DAgger (Dataset Aggregation)

Initialize $\mathcal{D} \leftarrow \emptyset$.
Initialize $\hat{\pi}_1$ to any policy in Π.
for $i = 1$ **to** N **do**
  Let $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$.
  Sample T-step trajectories using π_i.
  Get dataset $\mathcal{D}_i = \{(s, \pi^*(s))\}$ of visited states by π_i and actions given by expert.
  Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$.
  Train classifier $\hat{\pi}_{i+1}$ on $\mathcal{D}$.
end for
Return best $\hat{\pi}_i$ on validation.

π^*: expert's policy
$\beta_1 = 1, \beta_i \leq (1 - \alpha)^{i-1}$

Intuition: Building up the set of inputs that the learned policy is likely to encounter during its execution based on previous experience (training iterations).

- n_β: largest $n \leq N$ such that $\beta_n > \frac{1}{T}$
- $\epsilon_N = \min_{\pi \in \Pi} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{s \sim d_{\pi_i}} \left(l(s, \pi) \right)$; loss of the best policy in hindsight after N iterations
- $l_{\max}$: upper bound on the loss, i.e., $l_i(s, \hat{\pi}_i) \leq l_{\max}$ for all $\hat{\pi}_i$ and s

Lemma 4.1. $||d_{\pi_i} - d_{\hat{\pi}_i}||_1 \leq 2T\beta_i$.

Proof. Let d the distribution of states over T steps conditioned on π_i picking π^* at least once over T steps. Since π_i always executes $\hat{\pi}_i$ over T steps with probability $(1-\beta_i)^T$ we have $d_{\pi_i} = (1-\beta_i)^T d_{\hat{\pi}_i} + (1-(1-\beta_i)^T)d$. Thus

$$
\begin{aligned}
&||d_{\pi_i} - d_{\hat{\pi}_i}||_1 \\
&= (1-(1-\beta_i)^T)||d - d_{\hat{\pi}_i}||_1 \\
&\leq 2(1-(1-\beta_i)^T) \\
&\leq 2T\beta_i
\end{aligned}
$$

The last inequality follows from the fact that $(1-\beta)^T \geq 1 - \beta T$ for any $\beta \in [0, 1]$. □

DAgger

Theorem 4.1. *For* DAGGER, *there exists a policy* $\hat{\pi} \in \hat{\pi}_{1:N}$ *s.t.* $\mathbb{E}_{s\sim d_{\hat{\pi}}}[\ell(s,\hat{\pi})] \leq \epsilon_N + \gamma_N + \frac{2\ell_{\max}}{N}[n_\beta + T\sum_{i=n_\beta+1}^{N}\beta_i]$, *for* γ_N *the average regret of* $\hat{\pi}_{1:N}$.

Proof. The last lemma implies $\mathbb{E}_{s\sim d_{\hat{\pi}_i}}(\ell_i(s,\hat{\pi}_i)) \leq \mathbb{E}_{s\sim d_{\pi_i}}(\ell_i(s,\hat{\pi}_i)) + 2\ell_{\max}\min(1,T\beta_i)$. Then:

$$
\begin{aligned}
&\min_{\hat{\pi}\in\hat{\pi}_{1:N}} \mathbb{E}_{s\sim d_{\hat{\pi}}}[\ell(s,\hat{\pi})] \\
&\leq \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbb{E}_{s\sim d_{\hat{\pi}_i}}(\ell(s,\hat{\pi}_i)) \\
&\leq \tfrac{1}{N}\textstyle\sum_{i=1}^{N}[\mathbb{E}_{s\sim d_{\pi_i}}(\ell(s,\hat{\pi}_i)) + 2\ell_{\max}\min(1,T\beta_i)] \\
&\leq \gamma_N + \tfrac{2\ell_{\max}}{N}[n_\beta + T\textstyle\sum_{i=n_\beta+1}^{N}\beta_i] + \min_{\pi\in\Pi}\textstyle\sum_{i=1}^{N}\ell_i(\pi) \\
&= \gamma_N + \epsilon_N + \tfrac{2\ell_{\max}}{N}[n_\beta + T\textstyle\sum_{i=n_\beta+1}^{N}\beta_i]
\end{aligned}
$$

□

For $\beta_i \leq (1-\alpha)^{i-1}$,

$$\frac{1}{N}[n_\beta + T\sum_{i=n_\beta+1}^{N}\beta_i] \leq \frac{1}{N\alpha}[\log T + 1]$$

Theorem 3.1. *For* DAGGER, *if* N *is* $\tilde{O}(T)$ *there exists a policy* $\hat{\pi} \in \hat{\pi}_{1:N}$ *s.t.* $\mathbb{E}_{s\sim d_{\hat{\pi}}}[\ell(s,\hat{\pi})] \leq \epsilon_N + O(1/T)$

Theorem 3.2. *For* DAGGER, *if* N *is* $\tilde{O}(uT)$ *there exists a policy* $\hat{\pi} \in \hat{\pi}_{1:N}$ *s.t.* $J(\hat{\pi}) \leq J(\pi^*) + uT\epsilon_N + O(1)$.

cf. Supervised approach:

Theorem 2.1. *(Ross and Bagnell, 2010) Let* $\mathbb{E}_{s\sim d_{\pi^*}}[\ell(s,\pi)] = \epsilon$, *then* $J(\pi) \leq J(\pi^*) + T^2\epsilon$.