Efficient Reductions for Imitation Learning
Supplementary Material

Stéphane Ross
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA

J. Andrew Bagnell
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA

Detailed Proofs

Proof of Theorem 2.1

Let $\epsilon_i = \mathbb{E}_{s\sim d^i_{\pi^*}}[e_{\hat{\pi}}(s)]$ for $i = 1, 2, \ldots, T$ the expected 0-1 loss at time i of $\hat{\pi}$, such that $\epsilon = \frac{1}{T}\sum_{i=1}^T \epsilon_i$. Note that ϵ_t corresponds to the probability that $\hat{\pi}$ makes a mistake under distribution $d^t_{\pi^*}$. Let p_t represent the probability $\hat{\pi}$ hasn't made a mistake (w.r.t. π^*) in the first t-step, and d_t the distribution of state $\hat{\pi}$ is in at time t conditioned on the fact it hasn't made a mistake so far. If d'_t represents the distribution of states at time t obtained by following π^* but conditioned on the fact that $\hat{\pi}$ made at least one mistake in the first $t-1$ visited states. Then $d^t_{\pi^*} = p_{t-1}d_t + (1-p_{t-1})d'_t$. Now at time t, the expected cost of $\hat{\pi}$ is at most 1 if it has made a mistake so far, or $\mathbb{E}_{s\sim d_t}(C_{\hat{\pi}}(s))$ if it hasn't make a mistake yet. So $J(\hat{\pi}) \leq \sum_{t=1}^T [p_{t-1}\mathbb{E}_{s\sim d_t}(C_{\hat{\pi}}(s)) + (1-p_{t-1})]$. Let e_t and e'_t represent the probability of mistake of $\hat{\pi}$ in distribution d_t and d'_t. Then $\mathbb{E}_{s\sim d_t}(C_{\hat{\pi}}(s)) \leq \mathbb{E}_{s\sim d_t}(C_{\pi^*}(s)) + e_t$, and since $\epsilon_t = p_{t-1}e_t + (1-p_{t-1})e'_t$, then $p_{t-1}e_t \leq \epsilon_t$. Additionnally since $p_t = (1-e_t)p_{t-1}$, $p_t \geq p_{t-1} - \epsilon_t \geq 1 - \sum_{i=1}^t \epsilon_i$, i.e. $1 - p_t \leq \sum_{i=1}^t \epsilon_i$. Finally note that $J(\pi^*) = \sum_{t=1}^T [p_{t-1}\mathbb{E}_{s\sim d_t}(C_{\pi^*}(s)) + (1-p_{t-1})\mathbb{E}_{s\sim d'_t}(C_{\pi^*}(s))]$, so that $\sum_{t=1}^T p_{t-1}\mathbb{E}_{s\sim d_t}(C_{\pi^*}(s)) \leq J(\pi^*)$. Using these facts we obtain:

$$\begin{aligned}
J(\hat{\pi}) &\leq \textstyle\sum_{t=1}^T [p_{t-1}\mathbb{E}_{s\sim d_t}(C_{\hat{\pi}}(s)) + (1-p_{t-1})] \\
&\leq \textstyle\sum_{t=1}^T [p_{t-1}\mathbb{E}_{s\sim d_t}(C_{\pi^*}(s)) + p_{t-1}e_t + (1-p_{t-1})] \\
&\leq J(\pi^*) + \textstyle\sum_{t=1}^T \sum_{i=1}^t \epsilon_i \\
&\leq J(\pi^*) + T\textstyle\sum_{t=1}^T \epsilon_t \\
&= J(\pi^*) + T^2\epsilon
\end{aligned}$$

Proof of Theorem 3.1

At iteration i we are only changing the policy at step i, so $J(\pi^i) = J^{\pi^{i-1}}(\pi^i_i, i) = J(\pi^{i-1}) + \mathbb{A}(\pi^{i-1}, \pi^i_i)$. Solving this recurrence proves the theorem.

Proof of Lemma 4.1

With probability $\alpha^i(1-\alpha)^{T-i}\binom{T}{i}$, π^n executes π^{n-1} $T-i$ times and $\hat{\pi}^n$ i times. Hence $J(\pi^n) = (1-\alpha)^T J(\pi^{n-1}) + \sum_{i=1}^T \alpha^i(1-\alpha)^{T-i}\binom{T}{i}\bar{J}_i^{\pi^{n-1}}(\hat{\pi}^n)$. Since $(1-\alpha)^T = 1 - \sum_{i=1}^T \alpha^i(1-\alpha)^{T-i}\binom{T}{i}$, we obtain:

$$\begin{aligned}
J(\pi^n) &= J(\pi^{n-1}) + \textstyle\sum_{i=1}^T \alpha^i(1-\alpha)^{T-i}\binom{T}{i}\mathbb{A}_i(\pi^{n-1}, \hat{\pi}^n) \\
&\leq J(\pi^{n-1}) + \textstyle\sum_{i=1}^k \alpha^i(1-\alpha)^{T-i}\binom{T}{i}\mathbb{A}_i(\pi^{n-1}, \hat{\pi}^n) \\
&\quad + T\textstyle\sum_{i=k+1}^T \alpha^i(1-\alpha)^{T-i}\binom{T}{i}
\end{aligned}$$

for any k, since $\mathbb{A}_i(\pi, \pi') \leq T$ for any i, π, π'. The last summation term can be bounded as follows:

$$\begin{aligned}
&\textstyle\sum_{i=k}^T \alpha^i(1-\alpha)^{T-i}\binom{T}{i} \\
&= \textstyle\sum_{i=k}^T \alpha^i\binom{T}{i}\sum_{j=0}^{T-i}(-\alpha)^j\binom{T-i}{j} \\
&= \alpha^k \textstyle\sum_{i=0}^{T-k} \frac{T!i!}{(T-k)!(i+k)!}\sum_{j=0}^{T-k-i}\alpha^i(-\alpha)^j\frac{(T-k)!}{i!j!(T-k-i-j)!} \\
&\leq \alpha^k\binom{T}{k}\textstyle\sum_{i=0}^{T-k}\sum_{j=0}^{T-k-i}\alpha^i(-\alpha)^j\frac{(T-k)!}{i!j!(T-k-i-j)!} \\
&= \alpha^k\binom{T}{k}
\end{aligned}$$

where the inequality is true for $\alpha \leq \frac{1}{T}$ since $\sum_{j=0}^{T-k-i}(-\alpha)^j\frac{(T-k)!}{i!j!(T-k-i-j)!} \geq 0$ for such α and $\frac{T!i!}{(T-k)!(i+k)!} \leq \binom{T}{k}$ for all i. The first and last equality are from the binomial and multinomial theorem respectively. Hence we obtain a recurrence and expanding it up to $n = 0$ proves the lemma.

Proof of Lemma 4.2

With probability $(1-p_n)^T$, π^n never queries the expert and has T-step cost of $J(\tilde{\pi}^n)$, with probability $p_n(1-p_n)^{T-1}$, π^n queries once at time t and has T-step cost of $J_1^{\tilde{\pi}^n}(\pi^0, t)$ and in all other cases, it has cost ≥ 0. Since $(1-p_n)^T \geq (1-p_nT)$ we have that:

$$\begin{aligned}
J(\pi^n) &\geq (1-p_n)^T J(\tilde{\pi}^n) + p_nT(1-p_n)^{T-1}\bar{J}_1^{\tilde{\pi}^n}(\pi^0) \\
&\geq J(\tilde{\pi}^n) + p_nT[(1-p_n)^{T-1}\bar{J}_1^{\tilde{\pi}^n}(\pi^0) - J(\tilde{\pi}^n)] \\
&\geq J(\tilde{\pi}^n) - p_nT^2
\end{aligned}$$

Proof of Theorem 4.1

First, since for SMILe $\hat{\pi}^{n+1}$ will be close to π^n, we can derive bounds on the policy disadvantages. Let $\epsilon_{n+1} =$

$\mathbb{E}_{s\sim d_{\pi^n}}(e(s,\hat{\pi}^{*n+1}))$, then:

1. $\mathbb{A}_1(\pi^n,\hat{\pi}^{n+1}) = (1-\alpha)^n(\bar{J}_1^{\pi^n}(\hat{\pi}^{*n+1}) - \bar{J}_1^{\pi^n}(\pi^*))$
2. $\mathbb{A}_2(\pi^n,\hat{\pi}^{n+1}) \leq 2\mathbb{A}_1(\pi^n,\hat{\pi}^{n+1}) + 4(1-\alpha)^{2n}T\epsilon_{n+1}$

Proof of 1) This follows immediatly from the fact that $\bar{J}_1^{\pi^n}(\hat{\pi}^{n+1}) = \frac{1}{T}\sum_{t=1}^{T}[(1-\alpha)^n J_1^{\pi^n}(\hat{\pi}^{*n+1},t) + \alpha\sum_{i=1}^{n}(1-\alpha)^{i-1}J_1^{\pi^n}(\hat{\pi}^{*i},t)]$ and $J(\pi^n) = \frac{1}{T}\sum_{t=1}^{T}[(1-\alpha)^n J_1^{\pi^n}(\pi^*,t) + \alpha\sum_{i=1}^{n}(1-\alpha)^{i-1}J_1^{\pi^n}(\hat{\pi}^{*i},t)]$.

Proof of 2) Let $p_n = (1-\alpha)^n$. First notice that $J_1^{\pi}(\pi',t) = \frac{1}{T-1}\sum_{t'\neq t}J_2^{\pi}(\pi',\pi,t,t')$. Using this and the fact that $\hat{\pi}^{n+1} = p_n\hat{\pi}^{*n+1} + (1-p_n)\tilde{\pi}^n$ and $\pi^n = p_n\pi^* + (1-p_n)\tilde{\pi}^n$, we have that:

$$\begin{aligned}
&\mathbb{A}_1(\pi^n,\hat{\pi}^{n+1})\\
&= \tfrac{1}{T(T-1)}\textstyle\sum_{t=1}^{T-1}\sum_{t'=t+1}^{T} p_n^2[J_2^{\pi^n}(\hat{\pi}^{*n+1},\pi^*,t,t')\\
&\quad + J_2^{\pi^n}(\pi^*,\hat{\pi}^{*n+1},t,t') - 2J_2^{\pi^n}(\pi^*,t,t')]\\
&\quad + p_n(1-p_n)[J_2^{\pi^n}(\hat{\pi}^{*n+1},\tilde{\pi}^n,t,t') - J_2^{\pi^n}(\pi^*,\tilde{\pi}^n,t,t')]\\
&\quad + p_n(1-p_n)[J_2^{\pi^n}(\tilde{\pi}^n,\hat{\pi}^{*n+1},t,t') - J_2^{\pi^n}(\tilde{\pi}^n,\pi^*,t,t')]
\end{aligned}$$

Using this previous fact, we obtain that:

$$\begin{aligned}
&\mathbb{A}_2(\pi^n,\hat{\pi}^{n+1})\\
&= \tfrac{1}{\binom{T}{2}}\textstyle\sum_{t=1}^{T-1}\sum_{t'=t+1}^{T} p_n^2[J_2^{\pi^n}(\hat{\pi}^{*n+1},t,t') -\\
&\quad J_2^{\pi^n}(\pi^*,\hat{\pi}^{*n+1},t,t') + J_2^{\pi^n}(\pi^*,t,t') -\\
&\quad J_2^{\pi^n}(\hat{\pi}^{*n+1},\pi^*,t,t')] + 2\mathbb{A}_1(\pi^n,\hat{\pi}^{n+1})
\end{aligned}$$

The bound follows from the fact that when $\hat{\pi}^{*n+1}$ acts like π^* at timestep t, the term in brackets is 0, and when $\hat{\pi}^{*n+1}$ doesn't act like π^* at timestep t, it is less than $2T$.

Theorem 4.1 follows from these bound, Lemma 4.1 for $k=2$ and Lemma 4.2, choosing $\alpha = \frac{\sqrt{3}}{T^2\sqrt{\log T}}$ and $N = \frac{2}{\alpha}\log T$.

Proof of Lemma 4.3

To prove this, we will condition on the number of times k, that π^n executes $\tilde{\pi}^n$ (i.e. does not execute the experts policy). Since π^n does not execute the expert's policy k times over T steps with probability $(1-p_n)^k p_n^{T-k}\binom{T}{k}$, we have that: $D(\pi^n) = \sum_{k=0}^{T}(1-p_n)^k p_n^{T-k}\binom{T}{k}D_k^{\tilde{\pi}^n}(\tilde{\pi}^n)$. Now $\tilde{\pi}^n = \frac{1-p_{n-1}}{1-p_n}\tilde{\pi}^{n-1} + \frac{\alpha p_{n-1}}{1-p_n}\hat{\pi}^{*n}$. The theorem follows from the fact that if $\tilde{\pi}^n$ is executed k times, it will always execute $\tilde{\pi}^{n-1}$ over those k times with probability $(\frac{1-p_{n-1}}{1-p_n})^k$, and it will execute $\hat{\pi}^{*n}$ at least once with probability $1-(\frac{1-p_{n-1}}{1-p_n})^k$.

Example

The example in this section demonstrates that there exist problems where SMILe and Forward Training can guarantee strictly better performance than the traditional supervised approach, and where the traditional supervised approach achieves the $O(T^2\epsilon)$ regret bound.

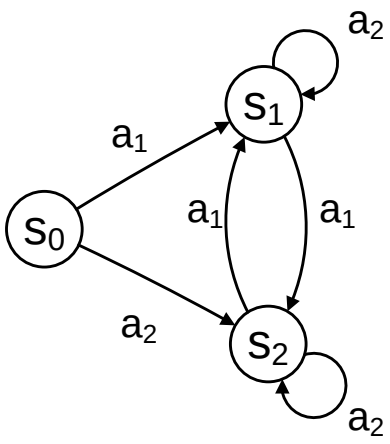

Figure 1: Problem where SMILe better than the supervised learning approach.

Consider the following problem with 3 states (s_0,s_1,s_2) and 2 actions (a_1,a_2). The agent always starts in s_0 and transitions are deterministic as specified in Figure 1.

The expert's policy π^* is to perform a_2 in s_1, and a_1 in s_0 and s_2, and consider the cost function we are trying to minimize is the imitation loss with respect to π^* (i.e. $C(s,a) = 1 - I(\pi^*(s),a)$, where I is the indicator function).

In this example, under π^*, one would only observe s_0 with frequency $\frac{1}{T}$ and s_1 the rest of the times, i.e. $d_{\pi^*} = (\frac{1}{T},\frac{T-1}{T},0)$. Now consider the policy $\hat{\pi}$ which executes a_1 with probability $(1-\epsilon T)$ in s_0, and a_2 in s_1, s_2, for some $\epsilon \leq \frac{1}{T}$. This policy which could be learned by the supervised learning approach achieves $\mathbb{E}_{s\sim d_{\pi^*}}(e(s,\hat{\pi})) = \epsilon$, however the T-step expected cost of $\hat{\pi}$ is $T^2\epsilon$ (with probability ϵT it has total cost of T, with probability $1-\epsilon T$ is has total cost of 0). This is an example where our upper bound in Theorem 2.1 is tight.

Now consider the Forward Training Algorithm. Here because the cost function is the imitation loss $u_i = 1$ for all i, as in any state, if we change the current action to perform π and then follow π^*, this will always have a total cost less than 1. Hence if $\epsilon_i = \mathbb{E}_{s\sim d^i_{\pi^{i-1}}}(e_{\pi^i_i}(s))$, then $\mathbb{A}(\pi^{i-1},\pi^i_i) = \epsilon_i$, so the forward training guarantees $T\bar{\epsilon}$ expected T-step cost on this problem, for $\bar{\epsilon} = \frac{1}{T}\sum_{i=1}^{T}\epsilon_i$.

Now consider the SMILe algorithm. Let $\hat{\pi}^{*n}$ denote the policy trained at iteration n under the state distribution $d_{\pi^{n-1}}$. In this problem, in any state, as soon as we do π^* we go to s_1. If we make a mistake in any state compared to executing π^*, we can only increase the T-step cost by 1 plus the expected number of steps it will take to come back to state s_1 under the current policy π^{n-1}. Since π^{n-1} executes π^* with probability at least $(1-\alpha)^{n-1}$, then this expected number of steps is at most $\frac{1}{(1-\alpha)^{n-1}}$. Hence for any policy π^{n-1},

$\sup_{s,\pi\in\Pi,t\leq T}[J_t^{\pi^{n-1}}(\pi,t,s) - J^{\pi^{n-1}}(\pi^*,t,s)] \leq 1 + \frac{1}{(1-\alpha)^{n-1}}$. Thus $\mathbb{A}(\pi^{n-1},\hat{\pi}^{*n}|\pi^*) \leq (1+\frac{1}{(1-\alpha)^{n-1}})\epsilon_n$, where $\epsilon_n = \mathbb{E}_{s\sim d_{\pi^{n-1}}}(e(s,\pi^{*n}))$. This gives us the following bound on $\tilde{\mathbb{A}}$:

$$\begin{aligned}\tilde{\mathbb{A}} &\leq \frac{\alpha}{1-(1-\alpha)^N}\sum_{i=1}^N (1-\alpha)^{i-1}(1+\frac{1}{(1-\alpha)^{i-1}})\epsilon_i \\ &\leq 2\frac{\alpha}{1-(1-\alpha)^N}\sum_{i=1}^N \epsilon_i \\ &= 2\frac{\alpha}{1-(1-\alpha)^N}N\bar{\epsilon}\end{aligned}$$

Also note that:

$$\begin{aligned}\tilde{\epsilon} &= \frac{\alpha}{1-(1-\alpha)^N}\sum_{i=1}^N (1-\alpha)^{i-1}\epsilon_i \\ &\leq \frac{\alpha}{1-(1-\alpha)^N}\sum_{i=1}^N \epsilon_i \\ &= \frac{\alpha}{1-(1-\alpha)^N}N\bar{\epsilon}\end{aligned}$$

Thus for $N = \frac{2}{\alpha}\ln T$, we have $\tilde{\mathbb{A}} \leq \frac{4}{1-\frac{1}{T^2}}\ln T\bar{\epsilon}$ and $\tilde{\epsilon} \leq \frac{2}{1-\frac{1}{T^2}}\ln T\bar{\epsilon}$. Hence SMILe guarantees an expected T-step cost of $O(T\log T\bar{\epsilon})$ on this example, which is better than the traditional supervised approach, but slightly worse than the forward training algorithm.