Agnostic System Identification for Model-Based Reinforcement Learning Supplementary Material

Stephane Ross STEPHANEROSS@CMU.EDU
Robotics Institute, Carnegie Mellon University, PA USA

J. Andrew Bagnell DBAGNELL@RI.CMU.EDU
Robotics Institute, Carnegie Mellon University, PA USA

1. Appendix - Detailed Analysis and Proofs

This supplementary material contains the detailed proofs and analysis of the theoretical results presented in the paper.

Additional Notation: We first introduce additional notation not used in the paper that is useful in some proofs. In particular, we define $d^t_{\omega,\pi}$ the distribution of states at time t if we executed π from time step 1 to $t-1$, starting from distribution ω at time 1, and $d_{\omega,\pi} = (1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}d^t_{\omega,\pi}$ the discounted distribution of states over the infinite horizon if we follow π, starting in ω at time 1.

1.1. Relating Performance to Error in Model

This subsection presents a number of useful lemmas for relating the performance (in terms of expected total cost) of a policy in the real system to the predictive error in the learned model from which the policy was computed.

Lemma 1.1. *Suppose we learned an approximate model $\hat{T}$ instead of the true model T and let $\hat{V}^\pi$ represent the value function of π under $\hat{T}$. Then for any state distribution ω:*

$$\begin{aligned}&\mathbb{E}_{s\sim\omega}[V^\pi(s) - \hat{V}^\pi(s)]\\ =\ & \tfrac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\pi}}[\mathbb{E}_{s'\sim T_{sa}}[\hat{V}^\pi(s')] - \mathbb{E}_{s'\sim\hat{T}_{sa}}[\hat{V}^\pi(s')]]\end{aligned}$$

Proof.

$$\begin{aligned}&\mathbb{E}_{s\sim\omega}[V^\pi(s) - \hat{V}^\pi(s)]\\ =\ & \mathbb{E}_{s\sim\omega,a\sim\pi_s}[C(s,a) + \gamma\mathbb{E}_{s'\sim T_{sa}}[V^\pi(s')]\\ & -C(s,a) - \gamma\mathbb{E}_{s'\sim\hat{T}_{sa}}[\hat{V}^\pi(s')]]\\ =\ & \gamma\mathbb{E}_{s\sim\omega,a\sim\pi_s}[\mathbb{E}_{s'\sim T_{sa}}[V^\pi(s')] - \mathbb{E}_{s'\sim\hat{T}_{sa}}[\hat{V}^\pi(s')]]\\ =\ & \gamma\mathbb{E}_{s\sim\omega,a\sim\pi_s}[\mathbb{E}_{s'\sim T_{sa}}[V^\pi(s')] - \mathbb{E}_{s'\sim T_{sa}}[\hat{V}^\pi(s')]\\ & +\mathbb{E}_{s'\sim T_{sa}}[\hat{V}^\pi(s')] - \mathbb{E}_{s'\sim\hat{T}_{sa}}[\hat{V}^\pi(s')]]\\ =\ & \gamma\mathbb{E}_{s\sim d^2_{\omega,\pi}}[V^\pi(s) - \hat{V}^\pi(s)]\\ & +\gamma\mathbb{E}_{(s,a)\sim D^1_{\omega,\pi}}[\mathbb{E}_{s'\sim T_{sa}}[\hat{V}^\pi(s')] - \mathbb{E}_{s'\sim\hat{T}_{sa}}[\hat{V}^\pi(s')]]\end{aligned}$$

This gives us a recurence. Solving this recurence proves the lemma. □

Corollary 1.1. *Suppose for all s,a: $C(s,a) \in [C_{\min}, C_{\max}]$, or for all s: $\hat{V}^\pi(s) \in [\hat{V}_{\min}, \hat{V}_{\max}]$, then:*

$$\begin{aligned}&\mathbb{E}_{s\sim\omega}[V^\pi(s) - \hat{V}^\pi(s)]\\ \leq\ & \tfrac{\gamma(\hat{V}_{\max}-\hat{V}_{\min})}{2(1-\gamma)}||\mathbb{E}_{(s,a)\sim D_{\omega,\pi}}[T_{sa} - \hat{T}_{sa}]||_1\\ \leq\ & \tfrac{\gamma(C_{\max}-C_{\min})}{2(1-\gamma)^2}\mathbb{E}_{(s,a)\sim D_{\omega,\pi}}[||T_{sa} - \hat{T}_{sa}||_1]\end{aligned}$$

Proof. Let $\Delta T = \mathbb{E}_{(s,a)\sim D_{\omega,\pi}}[T_{sa} - \hat{T}_{sa}]$. Note that $\sum_{s'}\Delta T(s') = 0$, so that for any constant $c \in \mathbb{R}$, $\sum_{s'} c\Delta T(s') = 0$. Then by the previous lemma we have that for any constant $c \in \mathbb{R}$:

$$\begin{aligned}&\mathbb{E}_{s\sim\omega}[V^\pi(s) - \hat{V}^\pi(s)]\\ =\ & \tfrac{\gamma}{1-\gamma}\textstyle\sum_{s'}\Delta T(s')\hat{V}_\pi(s')\\ =\ & \tfrac{\gamma}{1-\gamma}\textstyle\sum_{s'}\Delta T(s')(\hat{V}_\pi(s') - c)\\ \leq\ & \tfrac{\gamma}{1-\gamma}||\Delta T||_1 \sup_s|\hat{V}_\pi(s) - c|\end{aligned}$$

In particular, if $\hat{V}_\pi(s) \in [\hat{V}_{\min}, \hat{V}_{\max}]$ for all s, we can choose $c = \frac{\hat{V}_{\max}-\hat{V}_{\min}}{2}$ to guarantee that $\sup_s|\hat{V}_\pi(s) - c| \leq \frac{\hat{V}_{\max}-\hat{V}_{\min}}{2}$. Thus $\mathbb{E}_{s\sim\omega}[V^\pi(s) - \hat{V}^\pi(s)] \leq \frac{\gamma(\hat{V}_{\max}-\hat{V}_{\min})}{2(1-\gamma)}||\Delta T||_1$. If $C(s,a) \in [C_{\min}, C_{\max}]$ for all (s,a), then this implies $\hat{V}_\pi(s) \in [\hat{V}_{\min}, \hat{V}_{\max}]$ for all s for $\hat{V}_{\min} = \frac{C_{\min}}{1-\gamma}$ and $\hat{V}_{\max} = \frac{C_{\max}}{1-\gamma}$. Plugin in those values for $\hat{V}_{\min}$ and $\hat{V}_{\max}$, and the fact that $||\cdot||_1$ is convex with Jensen's inequality, proves the second result. The proof also applies in the continuous case by replacing the sum over s' by an integral over the state space in the first and second equality. □

Lemma 1.2. *Suppose we learned an approximate model $\hat{T}$ instead of the true model T and let $\hat{V}^\pi$ represent the value function of π under $\hat{T}$. Then for any state distribution ω*

and policies π, π':

$$\begin{aligned} & J_\omega(\pi) - J_\omega(\pi') \\ = \;& \mathbb{E}_{s\sim\omega}[\hat{V}^{\pi}(s) - \hat{V}^{\pi'}(s)] \\ & + \tfrac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\pi}}[\mathbb{E}_{s'\sim T_{sa}}[\hat{V}^{\pi}(s')] - \mathbb{E}_{s'\sim \hat{T}_{sa}}[\hat{V}^{\pi}(s')]] \\ & + \tfrac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\omega,\pi'}}[\mathbb{E}_{s'\sim \hat{T}_{sa}}[\hat{V}^{\pi'}(s')] - \mathbb{E}_{s'\sim T_{sa}}[\hat{V}^{\pi'}(s')]] \end{aligned}$$

Proof.

$$\begin{aligned} & J_\omega(\pi) - J_\omega(\pi') \\ = \;& \mathbb{E}_{s\sim\omega}[V^{\pi}(s) - V^{\pi'}(s)] \\ = \;& \mathbb{E}_{s\sim\omega}[(\hat{V}^{\pi}(s) - \hat{V}^{\pi'}(s)) + (V^{\pi}(s) - \hat{V}^{\pi}(s)) \\ & -(V^{\pi'}(s) - \hat{V}^{\pi'}(s))] \end{aligned}$$

Applying lemma 1.1 to $\mathbb{E}_{s\sim\omega}[V^{\pi}(s) - \hat{V}^{\pi}(s)]$ and $-\mathbb{E}_{s\sim\omega}[V^{\pi'}(s) - \hat{V}^{\pi'}(s)]$ proves the lemma. □

Suppose that $C(s,a) \in [C_{\min}, C_{\max}]$ for all s, a and let $C_{\text{rng}} = C_{\max} - C_{\min}$ and $H = \frac{\gamma C_{\text{rng}}}{(1-\gamma)^2}$.

Corollary 1.2. *Suppose we learned an approximate model $\hat{T}$ and solved it approximately to obtain π. For any policy π', let $\epsilon_{oc}^{\pi'} = \mathbb{E}_{s\sim\omega}[\hat{V}^{\pi}(s) - \hat{V}^{\pi'}(s)]$ denote how much larger is the expected total cost of π in the learned model $\hat{T}$ compared to π' for start distribution ω. Then for any policy π':*

$$J_\omega(\pi) - J_\omega(\pi') \leq \epsilon_{oc}^{\pi'} + H\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1]$$

for $D = \frac{1}{2}D_{\omega,\pi} + \frac{1}{2}D_{\omega,\pi'}$

Proof. Using lemma 1.2, we first note that the term $\mathbb{E}_{s\sim\omega}[\hat{V}^{\pi}(s) - \hat{V}^{\pi'}(s)] = \epsilon_{\text{oc}}^{\pi'}$. The other two terms can be bounded by $\frac{1}{2}H\mathbb{E}_{(s,a)\sim D_{\omega,\pi}}[||T_{sa} - \hat{T}_{sa}||_1]$ and $\frac{1}{2}H\mathbb{E}_{(s,a)\sim D_{\omega,\pi'}}[||T_{sa} - \hat{T}_{sa}||_1]$ respectively, following similar steps as in the proof of corollary 1.1. Combining those two terms proves the corollary. □

This corollary forms the basis of much of our analysis of the *Batch* and DAgger algorithms. In fact, this corollary already provides a performance bound for *Batch*, albeit with a major caveat: it bounds test performance of the learned policy π as a function of an error notion in the learned model $\hat{T}$ that cannot be minimized or controlled by the algorithm. That is, when collecting data under exploration distribution ν and fitting the model $\hat{T}$ based on this data, *Batch* could be making the quantity $\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1]$ arbitrarily close to its maxima (i.e. 2) in order to achieve low expected error under the training distribution ν. Even if there exists a model $T' \in \mathcal{T}$ where $\mathbb{E}_{(s,a)\sim D}[||T_{sa} - T'_{sa}||_1]$ is small, *Batch* would not pick this model if it has larger error under ν compared to other models in the class $\mathcal{T}$. As is, this bound only says that: if by chance *Batch* ends up picking a model that has low error under distribution D, then it must find a policy π not much worse than π'. Instead we would like to be able to say something much stronger of the form: if there exists a model with low error on training data, then we must find a policy that performs well compared to other policies π'. To do so, we must bound the term $\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1]$ by a training error term that the algorithm is minimizing. A first issue is bounding the L_1 distance by a loss we can minimize from observed samples. We present several possibilities for this in the next section. Then the remaining part will simply involve performing a change of distribution to bound the error under distribution D in terms of the error under the training distribution.

1.2. Relating L_1 distance to observable losses

This subsection presents a number of useful lemmas for relating the predictive error in L_1 distance that we would ideally need to minimize to other losses that are easier to minimize when learning a model from sampled transitions. These results prove Lemma 3.1 in the paper.

1.2.1. Relation to Classification Loss

We first show how the L_1 distance can be related to a classification loss when learning deterministic transition models in MDPs with finitely many states. Namely, given a model $\hat{T}$ which predicts next state $\hat{s}'_{sa}$ when doing action a in state s, then we define the 0-1 classification loss of $\hat{T}$ when observing transition (s, a, s') as:

$$\ell_{0-1}(\hat{T}, s, a, s') = I(s' \neq \hat{s}'_{sa}),$$

for I the indicator function. We show below that the L_1 distance is related to this classification loss by the following:

$$\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1] = 2\mathbb{E}_{(s,a)\sim D, s'\sim T_{sa}}[\ell_{0-1}(\hat{T}, s, a, s')]$$

This is proven in the following lemma:

Lemma 1.3. *Suppose $\hat{T}$ is a deterministic transition function (i.e. for any s, a, $\hat{T}_{sa}$ has probability 1 on a particular next state $\hat{s}'_{sa}$), e.g. a multiclass classifier. Then for any joint state-action distribution D, $\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1] = 2\mathbb{E}_{(s,a)\sim D, s'\sim T_{sa}}[\ell_{0-1}(\hat{T}, s, a, s')]$.*

Proof.

$$\begin{aligned} & \mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1] \\ = \;& \mathbb{E}_{(s,a)\sim D}[\textstyle\sum_{s'} |T_{sa}(s') - \hat{T}_{sa}(s')|] \\ = \;& \mathbb{E}_{(s,a)\sim D}[1 - T_{sa}(\hat{s}'_{sa}) + \textstyle\sum_{s'\neq\hat{s}'_{sa}} T_{sa}(s')] \\ = \;& 2\mathbb{E}_{(s,a)\sim D}[P_{s'\sim T_{sa}}(s' \neq \hat{s}'_{sa})] \\ = \;& 2\mathbb{E}_{(s,a)\sim D, s'\sim T_{sa}}[I(s' \neq \hat{s}'_{sa})] \end{aligned}$$

□

Additionally, any surrogate loss ℓ that upper bounds the 0-1 loss that are often used when learning classifiers (e.g. hinge loss when learning SVMs) could be used to upper bound

the L_1 distance. In this case, we have $\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1] \leq 2\mathbb{E}_{(s,a)\sim D, s'\sim T_{sa}}[\ell(\hat{T}, s, a, s')]$ from the fact that $\ell(\hat{T}, s, a, s') \geq \ell_{0-1}(\hat{T}, s, a, s')$. This proves the statement $\epsilon^{\text{L1}}_{\text{prd}} \leq 2\epsilon^{\text{cls}}_{\text{prd}}$ in Lemma 3.1 of the paper.

1.2.2. Relation to Negative Log Likelihood

We now show that for arbitrary MDPs and set of models, we can minimize the negative log likelihood to minimize a bound on the L_1 distance. Namely, for any model $\hat{T}$, define the negative log likelihood loss on transition (s, a, s') as:

$$\ell_{\text{nlh}}(\hat{T}, s, a, s') = -\log(\hat{T}_{sa}(s')).$$

Then this loss can be related to the L_1 distance as follows:

$$\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1] \leq \sqrt{2\mathbb{E}_{(s,a)\sim D, s'\sim T_{sa}}[\ell_{\text{nlh}}(\hat{T}, s, a, s') - \ell_{\text{nlh}}(T, s, a, s')]}.$$

This is shown in the lemma below:

Lemma 1.4. *For any joint state-action distribution D, $\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1] \leq \sqrt{2\mathbb{E}_{(s,a)\sim D, s'\sim T_{sa}}[\ell_{nlh}(\hat{T}, s, a, s') - \ell_{nlh}(T, s, a, s')]}$.*

Proof. We know that $||T_{sa} - \hat{T}_{sa}||_1 = 2||T_{sa} - \hat{T}_{sa}||_{\text{tv}}$ for $||T_{sa} - \hat{T}_{sa}||_{\text{tv}}$ the total variation distance between T_{sa} and $\hat{T}_{sa}$. Additionally, Pinsker's inequality tells us that $||T_{sa} - \hat{T}_{sa}||_{\text{tv}} \leq \sqrt{\frac{\text{KL}(T_{sa}||\hat{T}_{sa})}{2}}$ for $\text{KL}(T_{sa}||\hat{T}_{sa}) = \mathbb{E}_{s'\sim T_{sa}}[\log(\frac{T_{sa}(s')}{\hat{T}_{sa}(s')})]$ the Kullback-Leibler divergence. Thus we have $||T_{sa} - \hat{T}_{sa}||_1 \leq \sqrt{2\,\text{KL}(T_{sa}||\hat{T}_{sa})}$. Hence:

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1] \\
&\leq \mathbb{E}_{(s,a)\sim D}[\sqrt{2\,\text{KL}(T_{sa}||\hat{T}_{sa})}] \\
&\leq \sqrt{2\mathbb{E}_{(s,a)\sim D}[\text{KL}(T_{sa}||\hat{T}_{sa})]} \\
&= \sqrt{2\mathbb{E}_{(s,a)\sim D, s'\sim T_{sa}}[\ell_{\text{nlh}}(\hat{T}, s, a, s') - \ell_{\text{nlh}}(T, s, a, s')]}
\end{aligned}$$

where the second inequality follows from the Jensen's inequality since $\sqrt{\cdot}$ is concave. □

This proves the statement $\epsilon^{\text{L1}}_{\text{prd}} \leq \sqrt{2\epsilon^{\text{KL}}_{\text{prd}}}$ in Lemma 3.1of the paper.

1.2.3. Relation to Squared Loss in the Mean

Another interesting special case not discussed in the paper is for continuous MDPs with additive gaussian noise and known covariance matrix where we seek to learn to predict the mean next state. In this case, we can relate the L_1 distance to a squared loss in predicting the mean next state. Namely, suppose that for all s, a, T_{sa} and $\hat{T}_{sa}$ are gaussian distributions, both with covariance matrix $\Sigma \succ 0$. Let μ_{sa} and $\hat{\mu}_{sa}$ denote the mean of T_{sa} and $\hat{T}_{sa}$ respectively. We define the squared loss of $\hat{T}$ on transition (s, a, s') as:

$$\ell_{\text{sq}}(\hat{T}, s, a, s') = ||\hat{\mu}_{sa} - s'||_2^2.$$

This loss can be related to the L_1 distance between T_{sa} and $\hat{T}_{sa}$ as follows:

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1] \\
&\leq c\sqrt{\mathbb{E}_{(s,a)\sim D, s'\sim T_{sa}}[\ell_{\text{sq}}(\hat{T}, s, a, s') - \ell_{\text{sq}}(T, s, a, s')]},
\end{aligned}$$

for $c = \sqrt{\frac{2}{\pi\sigma_{\min}(\Sigma)}}$ and $\sigma_{\min}(\Sigma)$ the minimum singular value of the noise covariance matrix Σ. This is proven in the two lemmas below:

Lemma 1.5. *Suppose X_1 and X_2 are 2 independent gaussian random variables such that $X_1 \sim N(\mu_1, \Sigma)$ and $X_2 \sim N(\mu_2, \Sigma)$ and denote G_1 and G_2 the pdf of X_1 and X_2. Then $||G_1 - G_2||_1 \leq \sqrt{\frac{2}{\pi\sigma_{\min}(\Sigma)}}||\mu_1 - \mu_2||_2$, for $\sigma_{\min}(A)$ the minimum singular value of matrix A.*

Proof. We have that $||G_1 - G_2||_1 = 2[P(X_1 \in A) - P(X_2 \in A)]$ for $A = \{x|G_1(x) \geq G_2(x)\}$. It can be seen that $G_1(x) \geq G_2(x)$ when $\theta^\top(x - \mu_1) \leq \tau$, for $\theta^\top = \frac{(\mu_2-\mu_1)^\top\Sigma^{-1}}{||\Sigma^{-1/2}(\mu_2-\mu_1)||_2}$, $\tau = \frac{||\Sigma^{-1/2}(\mu_2-\mu_1)||_2}{2}$ and $\Sigma^{-1/2}$ denote the matrix square root of Σ^{-1} (which exists since Σ^{-1} if symmetric positive definite). Thus $A = \{x|\theta^\top(x - \mu_1) \leq \tau\}$. Define random variables $Z_1 = \theta^\top(X_1 - \mu_1)$ and $Z_2 = \theta^\top(X_2 - \mu_1)$. Then we have that $Z_1 \sim N(0, 1)$ (i.e. a standard normal distribution) and $Z_2 \sim N(2\tau, 1)$. Thus:

$$\begin{aligned}
&||G_1 - G_2||_1 \\
&= 2[P(X_1 \in A) - P(X_2 \in A)] \\
&= 2[P(Z_1 \leq \tau) - P(Z_2 \leq \tau)] \\
&= 4\Phi(\tau) - 2
\end{aligned}$$

For Φ the cdf of a standard normal variable. Because $\tau \geq 0$ and $\Phi(x)$ is concave for $x \geq 0$, then we can upperbound $\Phi(\tau)$ with a first-order taylor series expansion about 0. Let ϕ denote the pdf of a standard normal distribution and $\sigma_{\max}(A)$ the maximum singular value of a matrix A, then we obtain:

$$\begin{aligned}
&4\Phi(\tau) - 2 \\
&\leq 4(\Phi(0) + \tau\phi(0)) - 2 \\
&= 4\tau\phi(0) \\
&= \sqrt{\tfrac{2}{\pi}}||\Sigma^{-1/2}(\mu_1 - \mu_2)||_2 \\
&\leq \sqrt{\tfrac{2}{\pi}}\sigma_{\max}(\Sigma^{-1/2})||\mu_1 - \mu_2||_2 \\
&= \sqrt{\tfrac{2}{\pi\sigma_{\min}(\Sigma)}}||\mu_1 - \mu_2||_2
\end{aligned}$$

□

Lemma 1.6. *Suppose that for all s, a, T_{sa} and $\hat{T}_{sa}$ are gaussian distributions, both with covariance matrix $\Sigma \succ 0$. Then for any joint state-action distribution D, $\mathbb{E}_{(s,a)\sim D}[||T_{sa} - \hat{T}_{sa}||_1] \leq$*

$\sqrt{\frac{2}{\pi\sigma_{\min}(\Sigma)}\mathbb{E}_{(s,a)\sim D,s'\sim T_{sa}}[\ell_{sq}(\hat{T},s,a,s')-\ell_{sq}(T,s,a,s')]}$, *for* $\sigma_{\min}(\Sigma)$ *the minimum singular value of matrix* Σ.

Proof. From Lemma 1.5, we directly have that $\mathbb{E}_{(s,a)\sim D}[||T_{sa}-\hat{T}_{sa}||_1] \leq \sqrt{\frac{2}{\pi\sigma_{\min}(\Sigma)}}\mathbb{E}_{(s,a)\sim D}[||\mu_{sa}-\hat{\mu}_{sa}||_2]$. Using the fact that $||\mu_{sa}-\hat{\mu}_{sa}||_2^2 = \mathbb{E}_{s'\sim T_{sa}}[||\hat{\mu}_{sa}-s'||_2^2 - ||\mu_{sa}-s'||_2^2]$ and that $\sqrt{\cdot}$ is concave with Jensen's inequality proves the lemma. □

1.3. Analysis of the Batch Algorithm

We now present the detailed analysis of the Batch Algorithm. As mentioned previously after corollary 1.2, this corollary already provides a performance bound for *Batch*, with the caveat that its performance is related to an error notion in the model that is not minimized by the algorihtm, and could be made arbitrarily large when *Batch* attemps to minimize error under the training distribution ν. As is, it only states that *Batch* gets good performance if by chance it picks a model with low error under the distribution $D = \frac{1}{2}D_{\omega,\hat{\pi}} + \frac{1}{2}D_{\omega,\pi'}$. To bound performance with respect to the model error *Batch* is minimizing, the proof will simply involve using Cor. 1.2, applying a change of distribution, as well as bounding the L_1 distance with alternate loss *Batch* can minimize from sample transitions using the results from the previous section.

Let's define $\epsilon_{\text{oc}}^{\pi'} = \mathbb{E}_{s\sim\mu}[\hat{V}^{\hat{\pi}}(s) - \hat{V}^{\pi'}(s)]$, for $\hat{V}^{\hat{\pi}}$ and $\hat{V}^{\pi'}$ the value functions of $\hat{\pi}$ and π' under learned model $\hat{T}$ respectively. The term $\epsilon_{\text{oc}}^{\pi'}$ measures how much better of a solution π' is compared to $\hat{\pi}$ (in terms of expected total cost) on the optimal control problem we solved (with the learned model $\hat{T}$). For instance, if we found an ϵ-optimal policy $\hat{\pi}$ within some class of policies Π for learned model $\hat{T}$, then $\epsilon_{\text{oc}}^{\pi'} \leq \epsilon$ for all $\pi' \in \Pi$. Define the predictive error of $\hat{T}$ on training distribution ν, measured in L_1 distance, as $\epsilon_{\text{prd}}^{\text{L1}} = \mathbb{E}_{(s,a)\sim\nu}[||T_{sa}-\hat{T}_{sa}||_1]$. Similarly, define $\epsilon_{\text{prd}}^{\text{KL}} = \mathbb{E}_{(s,a)\sim\nu,s'\sim T_{sa}}[\log(T_{sa}(s')) - \log(\hat{T}_{sa}(s'))]$ and $\epsilon_{\text{prd}}^{\text{cls}} = \mathbb{E}_{(s,a)\sim\nu,s'\sim T_{sa}}[\ell(\hat{T},s,a,s')]$ the training predictive error of $\hat{T}$ in terms of KL and classification loss respectively (ℓ is the 0-1 loss or any upper bound on the 0-1 loss such as hinge loss). Additionally, let $c_\nu^\pi = \sup_{s,a}[\frac{D_{\mu,\pi}(s,a)}{\nu(s,a)}]$ represent the mismatch between state-action distribution ν, and the state-action distribution induced by policy π starting in μ.

Theorem 3.1. *The policy* $\hat{\pi}$ *is s.t. for any policy* π'*:*

$$J_\mu(\hat{\pi}) \leq J_\mu(\pi') + \epsilon_{oc}^{\pi'} + \frac{c_\nu^{\hat{\pi}} + c_\nu^{\pi'}}{2} H\epsilon_{prd}^{L1}$$

Equivalently, using the relations in Section 1.2:

$$J_\mu(\hat{\pi}) \leq J_\mu(\pi') + \epsilon_{oc}^{\pi'} + \frac{c_\nu^{\hat{\pi}} + c_\nu^{\pi'}}{2} H\sqrt{2\epsilon_{prd}^{KL}}$$

$$J_\mu(\hat{\pi}) \leq J_\mu(\pi') + \epsilon_{oc}^{\pi'} + (c_\nu^{\hat{\pi}} + c_\nu^{\pi'})H\epsilon_{prd}^{cls}$$

Proof.

$$\begin{aligned}
& J_\mu(\hat{\pi}) - J_\mu(\pi') \\
\leq\ & \epsilon_{\text{oc}}^{\pi'} + \tfrac{H}{2}[\mathbb{E}_{(s,a)\sim D_{\mu,\hat{\pi}}}[||T_{sa}-\hat{T}_{sa}||_1] \\
& + \mathbb{E}_{(s,a)\sim D_{\mu,\pi'}}[||T_{sa}-\hat{T}_{sa}||_1]] \\
\leq\ & \epsilon_{\text{oc}}^{\pi'} + \tfrac{H}{2}[c_\nu^{\hat{\pi}}\mathbb{E}_{(s,a)\sim\nu}[||T_{sa}-\hat{T}_{sa}||_1] \\
& + c_\nu^{\pi'}\mathbb{E}_{(s,a)\sim\nu}[||T_{sa}-\hat{T}_{sa}||_1]] \\
=\ & \epsilon_{\text{oc}}^{\pi'} + \tfrac{c_\nu^{\hat{\pi}}+c_\nu^{\pi'}}{2} H\mathbb{E}_{(s,a)\sim\nu}[||T_{sa}-\hat{T}_{sa}||_1] \\
=\ & \epsilon_{\text{oc}}^{\pi'} + \tfrac{c_\nu^{\hat{\pi}}+c_\nu^{\pi'}}{2} H\epsilon_{\text{prd}}^{\text{L1}}
\end{aligned}$$

where the first inequality follows from corollary 1.2, and the second inequality follows from the fact that for any non-negative function f and distributions p, q, $\mathbb{E}_{x\sim p}[f(x)] \leq \sup_x[\frac{p(x)}{q(x)}]\mathbb{E}_{x\sim q}[f(x)]$. We now have proven the first statement of the theorem. Applying lemma 1.4 proves that $\epsilon_{\text{prd}}^{\text{L1}} \leq \sqrt{2\epsilon_{\text{prd}}^{\text{KL}}}$, from which the second statement follows. Similarly, lemma 1.3 proves that $\epsilon_{\text{prd}}^{\text{L1}} \leq 2\epsilon_{\text{prd}}^{\text{cls}}$, from which the third statement follows. □

This theorem relates performance of the learned policy $\hat{\pi}$ to the training error (under the exploration distribution ν) the algorithm is minimizing in the model fitting procedure. The factor $c_\nu^{\hat{\pi}}H$ represents by how much the error in the model $\hat{T}$ under training distribution ν can scale to larger errors in predicting total cost of the learned policy $\hat{\pi}$ in the real system T. Similarly $c_\nu^{\pi'}H$ represents by how much the error in the model $\hat{T}$ under training distribution ν can scale to larger errors in predicting total cost of another policy π' in the real system T. Together, with the error in solving the optimal control problem under $\hat{T}$, this bounds how much worse $\hat{\pi}$ can be compared to π'.

More interestingly, we can use this result to provide a strong guarantee of the form: if there exists a model in the class which achieves small enough error under the training distribution ν, *Batch* must find a policy with good test performance. We can guarantee this if we use consistent fitting procedures that converge to the best model in the class asymptotically, as we collect more and more data. This allows us to relate the predictive error to the capacity of the model class to achieve low predictive error under the training distribution ν. We denote the modeling error, measured in L_1 distance, as $\epsilon_{\text{mdl}}^{\text{L1}} = \inf_{T'\in\mathcal{T}}\mathbb{E}_{(s,a)\sim\nu}[||T_{sa}-T'_{sa}||_1]$. Similarly, define $\epsilon_{\text{mdl}}^{\text{KL}} = \inf_{T'\in\mathcal{T}}\mathbb{E}_{(s,a)\sim\nu,s'\sim T_{sa}}[\log(T_{sa}(s')) - \log(T'_{sa}(s'))]$ and $\epsilon_{\text{mdl}}^{\text{cls}} = \inf_{T'\in\mathcal{T}}\mathbb{E}_{(s,a)\sim\nu,s'\sim T_{sa}}[\ell(T',s,a,s')]$. These are all 0 in realizable settings, but generally non-zero in agnostic settings. After observing m sampled transitions, the generalization error $\epsilon_{\text{gen}}^{\text{L1}}(m,\delta)$ (or consistency rate) bounds with high probability $1-\delta$ the quantity $\epsilon_{\text{prd}}^{\text{L1}} - \epsilon_{\text{mdl}}^{\text{L1}}$. Similarly, $\epsilon_{\text{gen}}^{\text{KL}}(m,\delta)$ and $\epsilon_{\text{gen}}^{\text{cls}}(m,\delta)$ denote the generalization

error for the KL and classification loss respectively.

By definition, all these quantities are such that after observing m samples, with probability at least $1-\delta$: $\epsilon^{\text{L1}}_{\text{prd}} \leq \epsilon^{\text{L1}}_{\text{mdl}} + \epsilon^{\text{L1}}_{\text{gen}}(m,\delta)$, $\epsilon^{\text{KL}}_{\text{prd}} \leq \epsilon^{\text{KL}}_{\text{mdl}} + \epsilon^{\text{KL}}_{\text{gen}}(m,\delta)$ and $\epsilon^{\text{cls}}_{\text{prd}} \leq \epsilon^{\text{cls}}_{\text{mdl}} + \epsilon^{\text{cls}}_{\text{gen}}(m,\delta)$. If the procedure is consistent in minimizing the L_1 distance, this means $\epsilon^{\text{L1}}_{\text{gen}}(m,\delta) \to 0$ as $m \to \infty$ for any $\delta > 0$. Similarly, $\epsilon^{\text{KL}}_{\text{gen}}(m,\delta) \to 0$ and $\epsilon^{\text{cls}}_{\text{gen}}(m,\delta) \to 0$ as $m \to \infty$ for any $\delta > 0$ if the procedure is consistent in minimizing the KL and classification loss respectively. Combining with the previous result, this proves the following:

Corollary 3.1. *After observing m transitions, with probability at least $1-\delta$, for any policy π':*

$$J_\mu(\hat{\pi}) \leq J_\mu(\pi') + \epsilon_{oc}^{\pi'} + \frac{c_\nu^{\hat{\pi}} + c_\nu^{\pi'}}{2} H[\epsilon_{mdl}^{L1} + \epsilon_{gen}^{L1}(m,\delta)].$$

Equivalently, using the relations in Section 1.2:

$$\begin{aligned} J_\mu(\hat{\pi}) \quad \leq \quad & J_\mu(\pi') + \epsilon_{oc}^{\pi'} \\ & + \tfrac{c_\nu^{\hat{\pi}} + c_\nu^{\pi'}}{2} H \sqrt{2[\epsilon_{mdl}^{KL} + \epsilon_{gen}^{KL}(m,\delta)]}. \end{aligned}$$

$$J_\mu(\hat{\pi}) \leq J_\mu(\pi') + \epsilon_{oc}^{\pi'} + (c_\nu^{\hat{\pi}} + c_\nu^{\pi'})[\epsilon_{mdl}^{cls} + \epsilon_{gen}^{cls}(m,\delta)].$$

Additionally, if the fitting procedure is consistent w.r.t. the L1 distance, KL or classification loss then $\epsilon_{gen}^{L1}(m,\delta) \to 0$, $\epsilon_{gen}^{KL}(m,\delta) \to 0$, or $\epsilon_{gen}^{cls}(m,\delta) \to 0$ respectively as $m \to \infty$ for any $\delta > 0$.

This corollary can be used to prove sample complexity results for *Batch*. For example, with the classification loss, one could immediately leverage existing generalization error results from the supervised learning literature to determine the quantity $\epsilon^{\text{cls}}_{\text{gen}}(m,\delta)$ based on the particular class of hypothesis $\mathcal{T}$. These results would, e.g., express $\epsilon^{\text{cls}}_{\text{gen}}(m,\delta)$ as a function of the VC dimension (or multi-class equivalent) of $\mathcal{T}$. In many cases, Hoeffding's inequality combined with covering number arguments and a union bound can be used to compute these generalization error terms.

1.3.1. The Batch Algorithm's Performance Bound is Tight

As mentioned in the paper, the previous performance bound for *Batch* in Theorem 3.1 is tight, in that we can construct examples where the bound is achieved to an arbitrarily small additive constant. We here present such an example.

Consider the real system to be a MDP with 3 states (s_1, s_2, s_3) and 2 actions (a_1, a_2). The initial state is always s_1 (i.e. $\mu = [1;0;0]$). Executing action a_1 in s_1 and s_2 transits to s_1 with probability 1. Executing action a_2 in s_1 transits to s_2 with probability 1 and executing a_2 in s_2 transits to s_2 with large probability $1-\epsilon$, and transits to s_3 with small probability ϵ. Doing any action in s_3 transits back to s_3 with probability 1. There is small cost $\delta > 0$ for executing any action in s_1 and large cost of $C > \delta(1+\frac{1-\gamma}{\gamma\epsilon})$ for doing any action in s_3. Doing action a_2 in s_2 has 0 cost, and action a_1 is s_2 has cost δ.

In this system, an optimal policy always executes a_1 in s_2 and can execute any action in s_1 and s_3. So let's consider an optimal policy π^* that is uniform over (a_1, a_2) in s_1 and s_3. It achieves expected total cost of $J_\mu(\pi^*) = \frac{\delta}{1-\gamma}$.

Now consider that we learned a model $\hat{T}$ which is the same as the real system, except that the learned model predicts that when executing a_2 in s_2 it transits to s_2 with probability 1. The optimal policy under the learned model is to execute a_2 in s_1 and s_2, and to execute any action in s_3. So let's consider the policy $\hat{\pi}$ which is uniform over (a_1, a_2) in s_3 and picks a_2 in both s_1 and s_2. The distribution $d_{\mu,\hat{\pi}}$ induced by this policy can be computed as $d_{\mu,\hat{\pi}} = (1-\gamma)(I-\gamma T^{\hat{\pi}})^{-1}\mu$ where I is $|S| \times |S|$ the identity matrix, $T^{\hat{\pi}}$ is the transition matrix induced by $\hat{\pi}$ (element (i,j) corresponds to probability of transitioning from state j to state i when executing $\hat{\pi}$ in state j), and μ the vector containing the initial state distribution. It can be seen that the distribution $d_{\mu,\hat{\pi}} = [1-\gamma; \frac{\gamma(1-\gamma)}{1-\gamma(1-\epsilon)}; \frac{\gamma^2\epsilon}{1-\gamma(1-\epsilon)}]$ and the performance of the learned policy $\hat{\pi}$ in the real system is $J_\mu(\hat{\pi}) = \delta + \frac{\gamma^2\epsilon C}{(1-\gamma)(1-\gamma(1-\epsilon))}$. So we have that $J_\mu(\hat{\pi}) - J_\mu(\pi^*) = \frac{\gamma^2\epsilon C}{(1-\gamma)(1-\gamma(1-\epsilon))} - \frac{\gamma\delta}{1-\gamma}$.

Suppose the exploration distribution ν is induced by executing the policy π_0, which picks actions uniformly randomly in s_1 and s_3, and picks a_2 with small probability $\alpha > 0$ in s_2 (a_1 with large probability $1-\alpha$ in s_2). It can be seen that $\nu = d_{\mu,\pi_0} = \frac{1}{(1-\gamma)(1+\gamma/2-\gamma\alpha(1-\epsilon))+\gamma^2\alpha\epsilon/2}[(1-\gamma)(1-\gamma\alpha(1-\epsilon)); \gamma(1-\gamma)/2; \gamma^2\alpha\epsilon/2]$. Because the L_1 distance between the real system and learned model is 0 for all state-action pairs, except 2ϵ for state-action pair (s_2, a_2), we obtain that the predictive error during training is $\mathbb{E}_{(s,a)\sim\nu}[||T_{sa} - \hat{T}_{sa}||_1] = \frac{\gamma(1-\gamma)\alpha\epsilon}{(1-\gamma)(1+\gamma/2-\gamma\alpha(1-\epsilon))+\gamma^2\alpha\epsilon/2}$, which becomes arbitrarily small as $\alpha \to 0$. Thus the learned model could likely be picked by a model fitting procedure in practice for small α. The learned model is also an optimal model among deterministic models, so if $\mathcal{T}$ contains only deterministic models, $\hat{T}$ would likely be picked.

Now we have that $c_\nu^{\hat{\pi}} = \frac{2c}{\alpha}$ for $c = \frac{(1-\gamma)(1+\gamma/2-\gamma\alpha(1-\epsilon))+\gamma^2\alpha\epsilon/2}{1-\gamma(1-\epsilon)}$. Similarly, we have $d_{\mu,\pi^*} = \frac{1}{(1-\gamma)(1+\gamma/2)}[1-\gamma; \frac{\gamma(1-\gamma)}{2}; 0]$ so that $c_\nu^{\pi^*} = \frac{(1-\gamma)(1+\gamma/2-\gamma\alpha(1-\epsilon))+\gamma^2\alpha\epsilon/2}{(1-\alpha)(1-\gamma)(1+\gamma/2)}$.

For this problem we have $C_{\max} = C$ and $C_{\min} = 0$. Also since $\mathbb{E}_{s\sim\mu}[\hat{V}^{\hat{\pi}}(s)] = \delta$ and $\mathbb{E}_{s\sim\mu}[\hat{V}^{\pi^*}(s)] = \frac{\delta}{1-\gamma}$ we have $\epsilon_{\text{oc}}^{\pi^*} = -\frac{\gamma\delta}{1-\gamma}$. So using these quantities, we obtain that our

bound says that:

$$\begin{aligned}
&J_\mu(\hat{\pi}) - J_\mu(\pi^*) \\
&\leq \quad \epsilon_{\text{oc}}^{\pi^*} + \frac{c_\nu^{\hat{\pi}} + c_\nu^{\pi^*}}{2} H \epsilon_{\text{prd}}^{\text{L1}} \\
&= \quad \frac{\gamma C}{2(1-\gamma)^2}(c_\nu^{\hat{\pi}} + c_\nu^{\pi^*}) \frac{\gamma(1-\gamma)\alpha\epsilon}{(1-\gamma)(1+\gamma/2-\gamma\alpha(1-\epsilon))+\gamma^2\alpha\epsilon/2} - \frac{\gamma\delta}{1-\gamma} \\
&= \quad \frac{\gamma^2\alpha\epsilon C}{2(1-\gamma)}\left[\frac{2}{\alpha(1-\gamma(1-\epsilon))} + \frac{1}{(1-\alpha)(1-\gamma)(1+\gamma/2)}\right] - \frac{\gamma\delta}{1-\gamma} \\
&= \quad \frac{\gamma^2\epsilon C}{(1-\gamma)(1-\gamma(1-\epsilon))}\left[1 + \frac{\alpha(1-\gamma(1-\epsilon))}{(1-\alpha)(1-\gamma)(2+\gamma)}\right] - \frac{\gamma\delta}{1-\gamma}
\end{aligned}$$

As mentioned previously, we know that $J_\mu(\hat{\pi}) - J_\mu(\pi^*) = \frac{\gamma^2\epsilon C}{(1-\gamma)(1-\gamma(1-\epsilon))} - \frac{\gamma\delta}{1-\gamma}$. We observe that we can pick α arbitrarily close to 0 in the example above so that in the limit, as α becomes closer to 0, the bound becomes the exact value of $J_\mu(\hat{\pi}) - J_\mu(\pi^*)$. This shows that there exists examples where our bound is tight to an arbitrarily small additive constant.

1.4. Analysis of the DAgger Algorithm

We now present the detailed analysis of the DAgger Algorithm. Let's define $\bar{\epsilon}_{\text{oc}}^{\pi'} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{s\sim\mu}[\hat{V}_i(s) - \hat{V}_i^{\pi'}(s)]$, for $\hat{V}_i$ and $\hat{V}_i^{\pi'}$ the value functions of π_i and π' under learned model $\hat{T}^i$ respectively. The term $\bar{\epsilon}_{\text{oc}}^{\pi'}$ measures how much better of a solution π' is on average compared to the policies $\pi_{1:N}$ (in terms of expected total cost) on the optimal control problems we solved (with the learned models $\hat{T}^{1:N}$). For instance, if at each iteration i we found an ϵ_i-optimal policy π_i within some class of policies Π on learned model $\hat{T}^i$, then $\bar{\epsilon}_{\text{oc}}^{\pi'} \leq \frac{1}{N}\sum_{i=1}^N \epsilon_i$ for all $\pi' \in \Pi$. Additionally, define the average predictive error of $\hat{T}^{1:N}$ over the training iterations, measured in L_1 distance, as $\bar{\epsilon}_{\text{prd}}^{\text{L1}} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{(s,a)\sim\rho_i}[||\hat{T}_{sa}^i - T_{sa}||_1]$ for $\rho_i = \frac{1}{2}D_{\mu,\pi_i} + \frac{1}{2}\nu$ the state-action distribution used at iteration i to collect data. Similarly define $\bar{\epsilon}_{\text{prd}}^{\text{KL}} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{(s,a)\sim\rho_i, s'\sim T_{sa}}[\log(T_{sa}(s')) - \log(\hat{T}_{sa}^i(s'))]$ and $\bar{\epsilon}_{\text{prd}}^{\text{cls}} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{(s,a)\sim\rho_i, s'\sim T_{sa}}[\ell(\hat{T}, s, a, s')]$ the average training predictive error of $\hat{T}^{1:N}$ measured in KL and classification loss respectively (ℓ is 0-1 loss or any upper bound on the 0-1 loss such as hinge loss).

Lemma 4.1. *The policies $\pi_{1:N}$ are s.t. for any policy π':*

$$J_\mu(\hat{\pi}) \leq J_\mu(\overline{\pi}) \leq J_\mu(\pi') + \bar{\epsilon}_{oc}^{\pi'} + c_\nu^{\pi'} H \bar{\epsilon}_{prd}^{Ll}$$

Equivalently, using the results from Section 1.2:

$$J_\mu(\hat{\pi}) \leq J_\mu(\overline{\pi}) \leq J_\mu(\pi') + \bar{\epsilon}_{oc}^{\pi'} + c_\nu^{\pi'} H \sqrt{2\bar{\epsilon}_{prd}^{KL}}$$
$$J_\mu(\hat{\pi}) \leq J_\mu(\overline{\pi}) \leq J_\mu(\pi') + \bar{\epsilon}_{oc}^{\pi'} + 2c_\nu^{\pi'} H \bar{\epsilon}_{prd}^{cls}$$

Proof.

$$\begin{aligned}
&\min_{\pi\in\pi_{1:N}} J_\mu(\pi) - J_\mu(\pi') \\
&\leq \quad \frac{1}{N}\sum_{i=1}^N [J_\mu(\pi_i) - J_\mu(\pi')] \\
&\leq \quad \bar{\epsilon}_{\text{oc}}^{\pi'} + \frac{H}{2}\frac{1}{N}\sum_{i=1}^N [\mathbb{E}_{(s,a)\sim D_{\mu,\pi_i}}[||T_{sa} - \hat{T}_{sa}^i||_1] \\
&\quad\quad + \mathbb{E}_{(s,a)\sim D_{\mu,\pi'}}[||T_{sa} - \hat{T}_{sa}^i||_1]] \\
&\leq \quad \bar{\epsilon}_{\text{oc}}^{\pi'} + \frac{H}{2}\frac{1}{N}\sum_{i=1}^N [\mathbb{E}_{(s,a)\sim D_{\mu,\pi_i}}[||T_{sa} - \hat{T}_{sa}^i||_1] \\
&\quad\quad + c_\nu^{\pi'}\mathbb{E}_{(s,a)\sim\nu}[||T_{sa} - \hat{T}_{sa}^i||_1]] \\
&\leq \quad \bar{\epsilon}_{\text{oc}}^{\pi'} + \frac{c_\nu^{\pi'} H}{2}\frac{1}{N}\sum_{i=1}^N [\mathbb{E}_{(s,a)\sim D_{\mu,\pi_i}}[||T_{sa} - \hat{T}_{sa}^i||_1] \\
&\quad\quad + \mathbb{E}_{(s,a)\sim\nu}[||T_{sa} - \hat{T}_{sa}^i||_1]] \\
&= \quad \bar{\epsilon}_{\text{oc}}^{\pi'} + c_\nu^{\pi'} H \frac{1}{N}\sum_{i=1}^N [\mathbb{E}_{(s,a)\sim\rho_i}[||T_{sa} - \hat{T}_{sa}^i||_1] \\
&= \quad \bar{\epsilon}_{\text{oc}}^{\pi'} + c_\nu^{\pi'} H \bar{\epsilon}_{\text{prd}}^{\text{L1}}
\end{aligned}$$

where the second inequality follows from applying corollary 1.2 to each term $J_\mu(\pi_i) - J_\mu(\pi')$. □

The last lemma relates the performance of DAgger to the training loss of the sequence of models picked over the iterations of training. However it is only an intermediate step, and as is, it is unclear why it is meaningful. In particular, it is unclear why the term $\bar{\epsilon}_{\text{prd}}^{\text{L1}}$ (or $\bar{\epsilon}_{\text{prd}}^{\text{KL}}$, $\bar{\epsilon}_{\text{prd}}^{\text{cls}}$) should be small as it corresponds to an average loss of the models on out-of-training samples. That is, $\hat{T}^i$ is trained based on data seen so far from the distributions $\rho_1, \rho_2, \ldots, \rho_{i-1}$, but then its loss is evaluated under the distribution ρ_i in the term $\mathbb{E}_{(s,a)\sim\rho_i}[||T_{sa} - \hat{T}_{sa}^i||_1]$ (or $\mathbb{E}_{(s,a)\sim\rho_i, s'\sim T_{sa}}[\log(T_{sa}(s')) - \log(\hat{T}_{sa}^i(s'))]$, $\mathbb{E}_{(s,a)\sim\rho_i, s'\sim T_{sa}}[\ell(\hat{T}^i, s, a, s')]$) contributing to $\bar{\epsilon}_{\text{prd}}^{\text{L1}}$ (or $\bar{\epsilon}_{\text{prd}}^{\text{KL}}$, $\bar{\epsilon}_{\text{prd}}^{\text{cls}}$). So as is, it could be that $\bar{\epsilon}_{\text{prd}}^{\text{L1}}$ (or $\bar{\epsilon}_{\text{prd}}^{\text{KL}}$, $\bar{\epsilon}_{\text{prd}}^{\text{cls}}$) is large even if we achieve low error on the aggregate dataset at each iteration when fitting each $\hat{T}^i$. However we can observe that the quantity $\bar{\epsilon}_{\text{prd}}^{\text{L1}}$ (or $\bar{\epsilon}_{\text{prd}}^{\text{KL}}$, $\bar{\epsilon}_{\text{prd}}^{\text{cls}}$) can be interpreted as the average loss of an online learner on a particular online learning problem. This is where the no-regret property is crucial and makes this result interesting: no-regret guarantees that $\bar{\epsilon}_{\text{prd}}^{\text{L1}}$ (or $\bar{\epsilon}_{\text{prd}}^{\text{KL}}$, $\bar{\epsilon}_{\text{prd}}^{\text{cls}}$) must be small relative to the error of the best model in hindsight. So the combination of no-regret, and existence of a model with low error on the aggregate dataset, implies that $\bar{\epsilon}_{\text{prd}}^{\text{L1}}$ (or $\bar{\epsilon}_{\text{prd}}^{\text{KL}}$, $\bar{\epsilon}_{\text{prd}}^{\text{cls}}$) must be small. This is emphasized in the following theorem that constitutes our main result for DAgger.

We denote the modeling error under the overal training distribution $\bar{\rho} = \frac{1}{N}\sum_{i=1}^N \rho_i$, measured in L_1 distance as $\bar{\epsilon}_{\text{mdl}}^{\text{L1}} = \inf_{T'\in\mathcal{T}} \mathbb{E}_{(s,a)\sim\bar{\rho}}[||T_{sa} - T'_{sa}||]$. Similarly, denote $\bar{\epsilon}_{\text{mdl}}^{\text{KL}} = \inf_{T'\in\mathcal{T}} \mathbb{E}_{(s,a)\sim\bar{\rho}, s'\sim T_{sa}}[\log(T_{sa}(s')) - \log(T'_{sa}(s'))]$ and $\bar{\epsilon}_{\text{mdl}}^{\text{cls}} = \inf_{T'\in\mathcal{T}} \mathbb{E}_{(s,a)\sim\bar{\rho}, s'\sim T_{sa}}[\ell(T', s, a, s')]$ the modeling error measured in terms of KL and classification loss. The modeling error represents the error of the best model in hindsight after the N iterations of training. To relate the predictive error to this modeling error when using no-regret algorithms, we first need

to express the predictive error in terms of an online learning loss on a particular online learning problem. For each iteration $i \in 1:N$, define the following loss functions: $L_i^{\text{L1}}(\hat{T}) = \mathbb{E}_{(s,a)\sim\rho_i}[||T_{sa} - \hat{T}_{sa}||_1]$, $L_i^{\text{KL}}(\hat{T}) = \mathbb{E}_{(s,a)\sim\rho_i, s'\sim T_{sa}}[-\log(\hat{T}_{sa}(s'))]$, and $L_i^{\text{cls}}(\hat{T}) = \mathbb{E}_{(s,a)\sim\rho_i, s'\sim T_{sa}}[\ell(\hat{T}, s, a, s')]$. Now it can be seen that $\bar{\epsilon}_{\text{prd}}^{\text{L1}} = \frac{1}{N}\sum_{i=1}^N L_i^{\text{L1}}(\hat{T}^i)$, $\bar{\epsilon}_{\text{mdl}}^{\text{L1}} = \inf_{T'\in\mathcal{T}} \frac{1}{N}\sum_{i=1}^N L_i^{\text{L1}}(T')$, $\bar{\epsilon}_{\text{prd}}^{\text{KL}} = \frac{1}{N}\sum_{i=1}^N L_i^{\text{KL}}(\hat{T}^i) - L_i^{\text{KL}}(T)$, $\bar{\epsilon}_{\text{mdl}}^{\text{KL}} = \inf_{T'\in\mathcal{T}} \frac{1}{N}\sum_{i=1}^N L_i^{\text{KL}}(T') - L_i^{\text{KL}}(T)$, $\bar{\epsilon}_{\text{prd}}^{\text{cls}} = \frac{1}{N}\sum_{i=1}^N L_i^{\text{cls}}(\hat{T}^i)$ and $\bar{\epsilon}_{\text{mdl}}^{\text{cls}} = \inf_{T'\in\mathcal{T}} \frac{1}{N}\sum_{i=1}^N L_i^{\text{cls}}(T')$. DAgger uses a no-regret algorithm on one of the sequence of loss function $L_{1:N}^{\text{L1}}$, $L_{1:N}^{\text{KL}}$ or $L_{1:N}^{\text{cls}}$. If for instance we use the no-regret algorithm on the sequence of loss $L_{1:N}^{\text{KL}}$, then this implies that $\bar{\epsilon}_{\text{prd}}^{\text{KL}} - \bar{\epsilon}_{\text{mdl}}^{\text{KL}} = \frac{1}{N}\sum_{i=1}^N L_i^{\text{KL}}(\hat{T}^i) - \inf_{T'\in\mathcal{T}} \frac{1}{N}\sum_{i=1}^N L_i^{\text{KL}}(T') \to 0$ as $N \to \infty$. If we define $\bar{\epsilon}_{\text{rgt}}^{\text{KL}}$ the average regret of the online learning algorithm after N iterations when using the KL loss, then we have $\bar{\epsilon}_{\text{prd}}^{\text{KL}} \leq \bar{\epsilon}_{\text{mdl}}^{\text{KL}} + \bar{\epsilon}_{\text{rgt}}^{\text{KL}}$ for $\bar{\epsilon}_{\text{rgt}}^{\text{KL}} \to 0$ as $N \to \infty$. Similarly, if we use the classification loss, a no-regret algorithm on the sequence of loss $L_{1:N}^{\text{cls}}$ implies that $\bar{\epsilon}_{\text{prd}}^{\text{cls}} - \bar{\epsilon}_{\text{mdl}}^{\text{cls}} = \frac{1}{N}\sum_{i=1}^N L_i^{\text{cls}}(\hat{T}^i) - \inf_{T'\in\mathcal{T}} \frac{1}{N}\sum_{i=1}^N L_i^{\text{cls}}(T') \to 0$ as $N \to \infty$. If we define $\bar{\epsilon}_{\text{rgt}}^{\text{cls}}$ the average regret of the online learning algorithm after N iterations when using the classification loss, then we have $\bar{\epsilon}_{\text{prd}}^{\text{cls}} \leq \bar{\epsilon}_{\text{mdl}}^{\text{cls}} + \bar{\epsilon}_{\text{rgt}}^{\text{cls}}$ for $\bar{\epsilon}_{\text{rgt}}^{\text{cls}} \to 0$ as $N \to \infty$. While the L_1 distance cannot be evaluated from samples, some statistical estimators can be no-regret on the sequence of loss $L_{1:N}^{\text{L1}}$ with high probability without explicitly trying to minimize this loss. This is the case in finite MDPs if we use the empirical estimator of the transition matrix T based on all data seen so far over the iterations (see section 1.5.1). If we have a such sequence of models $\hat{T}_{1:N}$ which is no-regret on the sequence of loss $L_{1:N}^{\text{L1}}$, then $\bar{\epsilon}_{\text{prd}}^{\text{L1}} - \bar{\epsilon}_{\text{mdl}}^{\text{L1}} = \frac{1}{N}\sum_{i=1}^N L_i^{\text{L1}}(\hat{T}^i) - \inf_{T'\in\mathcal{T}} \frac{1}{N}\sum_{i=1}^N L_i^{\text{L1}}(T') \to 0$ as $N \to \infty$. If we define $\bar{\epsilon}_{\text{rgt}}^{\text{L1}}$ the average regret of $\hat{T}_{1:N}$ after N iterations on the L_1 distance, then we have $\bar{\epsilon}_{\text{prd}}^{\text{L1}} \leq \bar{\epsilon}_{\text{mdl}}^{\text{L1}} + \bar{\epsilon}_{\text{rgt}}^{\text{L1}}$ for $\bar{\epsilon}_{\text{L1}}^{\text{cls}} \to 0$ as $N \to \infty$. Combining with the previous lemma, this proves our main result:

Theorem 4.1. *The policies $\pi_{1:N}$ are s.t. for any policy π':*

$$J_\mu(\hat{\pi}) \leq J_\mu(\overline{\pi}) \leq J_\mu(\pi') + \bar{\epsilon}_{oc}^{\pi'} + c_\nu^{\pi'} H[\bar{\epsilon}_{mdl}^{L1} + \bar{\epsilon}_{rgt}^{L1}]$$

Equivalently, using the results from Section 1.2:

$$J_\mu(\hat{\pi}) \leq J_\mu(\overline{\pi}) \leq J_\mu(\pi') + \bar{\epsilon}_{oc}^{\pi'} + c_\nu^{\pi'} H\sqrt{2[\bar{\epsilon}_{mdl}^{KL} + \bar{\epsilon}_{rgt}^{KL}]}$$

$$J_\mu(\hat{\pi}) \leq J_\mu(\overline{\pi}) \leq J_\mu(\pi') + \bar{\epsilon}_{oc}^{\pi'} + 2c_\nu^{\pi'} H[\bar{\epsilon}_{mdl}^{cls} + \bar{\epsilon}_{rgt}^{cls}]$$

Additionally, if the fitting procedure is no-regret w.r.t. $L_{1:N}^{L1}$, $L_{1:N}^{KL}$, or $L_{1:N}^{cls}$, then $\bar{\epsilon}_{rgt}^{L1} \to 0$, $\bar{\epsilon}_{rgt}^{KL} \to 0$, or $\bar{\epsilon}_{rgt}^{cls} \to 0$ respectively, as $N \to 0$.

In cases where the distributions D_{μ,π_n} converge to a small region in the space of distributions as $n \to \infty$ (which tend to occur in practice), we can also guarantee good performance if we pick the last policy π_N, for N large enough:

Lemma 4.2. *Suppose there exists a distribution D^* and some $\epsilon_{cnv}^* \geq 0$ such that for all i, $||D_{\mu,\pi_i} - D^*||_1 \leq \epsilon_{cnv}^* + \epsilon_{cnv}^i$ for some sequence $\{\epsilon_{cnv}^n\}_{i=1}^\infty$ that is $o(1)$. Then the last policy π_N produced by DAgger is such that:*

$$J_\mu(\pi_N) \leq J_\mu(\overline{\pi}) + \frac{C_{rng}}{2(1-\gamma)}[2\epsilon_{cnv}^* + \epsilon_{cnv}^N + \frac{1}{N}\sum_{i=1}^N \epsilon_{cnv}^i]$$

Thus:

$$\limsup_{N\to\infty} J_\mu(\pi_N) - J_\mu(\overline{\pi}) \leq \frac{C_{rng}}{1-\gamma}\epsilon_{cnv}^*$$

Proof. We have that $D_{\mu,\overline{\pi}} = \frac{1}{N}\sum_{i=1}^N D_{\mu,\pi_i}$. By our assumptions, $||D_{\mu,\pi_N} - D_{\mu,\overline{\pi}}||_1 \leq 2\epsilon_{\text{cnv}}^* + \epsilon_{\text{cnv}}^N + \frac{1}{N}\sum_{i=1}^N \epsilon_{\text{cnv}}^i$. Thus:

$$\begin{aligned} & J_\mu(\pi_N) \\ = \; & \tfrac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\mu,\pi_N}}[C(s,a)] \\ \leq \; & \tfrac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim D_{\mu,\overline{\pi}}}[C(s,a)] + \tfrac{C_{\text{rng}}}{2(1-\gamma)}||D_{\mu,\pi_N} - D_{\mu,\overline{\pi}}||_1 \\ \leq \; & J_\mu(\overline{\pi}) + \tfrac{C_{\text{rng}}}{2(1-\gamma)}[2\epsilon_{\text{cnv}}^* + \epsilon_{\text{cnv}}^N + \tfrac{1}{N}\sum_{i=1}^N \epsilon_{\text{cnv}}^i] \end{aligned}$$

where the first inequality follows from the fact that for any function f, constant c, and distributions p, q, $\mathbb{E}_{x\sim p}[f(x)] \leq \mathbb{E}_{x\sim q}[f(x)] + \sup_x |f(x) - c||p - q||_1$. Here since $C(s,a) \in [C_{\min}, C_{\max}]$, choosing $c = \frac{C_{\text{rng}}}{2}$ minimizes the term $\sup_{s,a}|C(s,a) - c|$. □

1.5. Finite Sample Analysis for DAgger in Particular Scenarios

This subsection presents sample complexity results to achieve near-optimal performance with DAgger in two particular scenarios.

1.5.1. Finite MDP with Empirical Estimator

Consider the real system to be an arbitrary finite MDP with $|S|$ states and $|A|$ actions, and the model $\hat{T}^i$ used at iteration i to be the empirical estimator of T from the observed transitions in the first $i-1$ iterations. That is let $n_{sas'}^i$ be the number of times we observed transition (s,a,s') at iteration i (i.e. when sampling s, a from distribution $\rho_i = \frac{1}{2}D_{\mu,\pi_i} + \frac{1}{2}\nu$). Let $n_{sas'}^{<i} = \sum_{k=1}^{i-1} n_{sas'}^k$ the total number of times we observed state transition (s,a,s') in the first $i-1$ iterations, and $n_{sa}^{<i} = \sum_{s'} n_{sas'}^{<i}$ the number of times we picked sampled transitions from state action pair (s,a) in the first $i-1$ iterations. Then the empirical estimator at iteration i is such that $\hat{T}_{sa}^i(s') = \frac{n_{sas'}^{<i}}{n_{sa}^{<i}}$. If $n_{sa}^{<i} = 0$, then simply define $\hat{T}_{sa}^i(s') = \frac{1}{|S|}$. We seek to bound $\bar{\epsilon}_{\text{prd}}^{\text{L1}}$ after N iterations with high probability when using this empirical estimator and sampling m transitions

at each iteration.

Let $(s_{ij}, a_{ij}, s'_{ij})$ denote the j^{th} transition sampled at iteration i. For $i \in \{1, 2, \ldots, N\}$ and $j \in \{1, 2, \ldots, m\}$, define the random variables $Y_{(i-1)m+j} = \mathbb{E}_{(s,a)\sim\rho_i}[||T_{sa} - \hat{T}^i_{sa}||_1] - ||T_{s_{ij}a_{ij}} - \hat{T}^i_{s_{ij}a_{ij}}||_1$. Then $\mathbb{E}[Y_{(i-1)m+j}|Y_1, Y_2, \ldots, Y_{(i-1)m+j-1}] = 0$. Thus the random variables $X_k = \sum_{l=1}^{k} Y_l$ for $k \in \{1, 2, \ldots, Nm\}$ form a martingale. Since $Y_l \in [-2, 2]$, then by the Azuma-Hoeffding inequality we have $\frac{X_{Nm}}{Nm} \leq 2\sqrt{\frac{2\log(1/\delta)}{Nm}}$ with probability at least $1 - \delta$. Hence we must have that with probability at least $1 - \delta$:

$$\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbb{E}_{(s,a)\sim\rho_i}||\hat{T}^i_{sa} - T_{sa}||_1 \\
\leq\ & \tfrac{1}{Nm}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{m} ||\hat{T}^i_{s_{ij}a_{ij}} - T_{s_{ij}a_{ij}}||_1 + 2\sqrt{\tfrac{2\log(1/\delta)}{Nm}} \\
=\ & \tfrac{1}{Nm}\textstyle\sum_{i=1}^{N}\sum_{s,a} n^i_{sa}||\hat{T}^i_{sa} - T_{sa}||_1 + 2\sqrt{\tfrac{2\log(1/\delta)}{Nm}}
\end{aligned}$$

By applying a result from Wasserman (2003), we know that if we have m samples from a distribution P over k events and $\hat{P}$ denotes the empirical estimate of this distribution, then with probability at least $1 - \delta'$, $||\hat{P} - P||_1 \leq \sqrt{\frac{2\ln(2)k + 2\log(1/\delta')}{m}}$. Using an union bound, we conclude that with probability at least $1 - \delta'$, we must have that for all state-action pair s, a and iteration i:

$$||\hat{T}^i_{sa} - T_{sa}||_1 \leq \sqrt{\frac{2\log(2)|S| + 2\log(|S||A|N/\delta')}{n^{<i}_{sa}}}$$

It is also clear that $||\hat{T}^i_{sa} - T_{sa}||_1 \leq 2$ always hold. Thus we must have that with probability at least $1 - \delta - \delta'$:

$$\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbb{E}_{(s,a)\sim\rho_i}||\hat{T}^i_{sa} - T_{sa}||_1 \\
\leq\ & \tfrac{1}{Nm}\textstyle\sum_{i=1}^{N}\sum_{s,a} n^i_{sa} \min(2, \sqrt{\tfrac{2\log(2)|S| + 2\log(|S||A|N/\delta')}{n^{<i}_{sa}}}) \\
&+2\sqrt{\tfrac{2\log(1/\delta)}{Nm}}
\end{aligned}$$

The term $\min(2, \sqrt{\frac{2\ln(2)|S| + 2\log(|S||A|N/\delta')}{n^{<i}_{sa}}}) = 2$ when $n^{<i}_{sa} \leq m_0$ for $m_0 = \frac{2\log(2)|S| + 2\log(|S||A|N/\delta')}{4}$. Let $k_{sa} \in \{1, 2, \ldots, N\}$ be the largest iteration such that $n^{<k_{sa}}_{sa} \leq m_0$. Then we have that for all s, a:

$$\begin{aligned}
&\textstyle\sum_{i=1}^{N} n^i_{sa} \min(2, \sqrt{\frac{2\ln(2)|S| + 2\ln|S||A|N/\delta'}{n^{<i}_{sa}}}) \\
=\ & 2\textstyle\sum_{i=1}^{k_{sa}} n^i_{sa} + 2\sqrt{m_0}\sum_{i=k_{sa}+1}^{N} \frac{n^i_{sa}}{\sqrt{n^{<i}_{sa}}} \\
\leq\ & 2(m_0 + m) + 2\sqrt{m_0}\textstyle\sum_{i=k_{sa}+1}^{N} \frac{n^i_{sa}}{\sqrt{m_0 + \sum_{j=k_{sa}+1}^{i-1} n^j_{sa}}}
\end{aligned}$$

Thus we obtain that with probability at least $1 - \delta - \delta'$:

$$\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbb{E}_{(s,a)\sim\rho_i}[||\hat{T}^i_{sa} - T_{sa}||_1] \\
\leq\ & \tfrac{2m_0|S||A|}{Nm} + \tfrac{2|S||A|}{N} + 2\sqrt{\tfrac{2\log(1/\delta)}{Nm}} \\
&+\tfrac{2\sqrt{m_0}}{Nm}\textstyle\sum_{s,a}\sum_{i=k_{sa}+1}^{N} \frac{n^i_{sa}}{\sqrt{m_0 + \sum_{j=k_{sa}+1}^{i-1} n^j_{sa}}}
\end{aligned}$$

To upper bound this term, we will seek to upper bound $\sum_{s,a}\sum_{i=k_{sa}+1}^{N} \frac{n^i_{sa}}{\sqrt{m_0 + \sum_{j=k_{sa}+1}^{i-1} n^j_{sa}}}$ with respect to any choice of $\{n^i_{sa}\}$ an adversary might pick under the constraint that $\sum_{s,a} n^j_{sa} = m$ for all j. We have that:

$$\begin{aligned}
&\max_{\{n^i_{sa}\}} \textstyle\sum_{s,a}\sum_{i=k_{sa}+1}^{N} \frac{n^i_{sa}}{\sqrt{m_0 + \sum_{j=k_{sa}+1}^{i-1} n^j_{sa}}} \\
\leq\ & \max_{\{n^i_{sa}\}} \textstyle\sum_{s,a}\sum_{i=1}^{N} \frac{n^i_{sa}}{\sqrt{m_0 + \sum_{j=1}^{i-1} n^j_{sa}}} \\
=\ & \max_{\{n^i_{sa}\}} \textstyle\sum_{s,a}\sum_{i=1}^{N} \frac{n^i_{sa}}{\sqrt{m_0 + n^{<i}_{sa}}}
\end{aligned}$$

The inequality holds because for any assignment of $\{n^i_{sa}\}$, we can create a new assignment $\{n'^i_{sa}\}$ such that $\sum_{s,a}\sum_{i=1}^{N} \frac{n'^i_{sa}}{\sqrt{m_0 + \sum_{j=1}^{i-1} n'^j_{sa}}} \geq \sum_{s,a}\sum_{i=k_{sa}+1}^{N} \frac{n^i_{sa}}{\sqrt{m_0 + \sum_{j=k_{sa}+1}^{i-1} n^j_{sa}}}$ (namely by setting $n'^i_{sa} = n^{k_{sa}+i}_{sa}$ for $i \in \{1, 2, \ldots, N - k_{sa}\}$ and n'^i_{sa} arbitrarily for $i > N - k_{sa}$ for all s, a).

Now, it can be seen that $\sum_{s,a}\sum_{i=1}^{N} \frac{n^i_{sa}}{\sqrt{m_0 + n^{<i}_{sa}}}$ is maximized by sequentially setting the n^i_{sa} equal to m to the pair (s, a) with smallest $n^{<i}_{sa}$ and $n^i_{s'a'} = 0$ for all other (s', a') (and breaking ties arbitrarily). This implies that for iteration i such that $k|S||A| \leq i < (k+1)|S||A|$ for some non-negative integer k, $\sum_{s,a} \frac{n^i_{sa}}{\sqrt{m_0 + n^{<i}_{sa}}} \leq \frac{m}{\sqrt{m_0 + km}}$. For any N, let us express $N = k|S||A| + l$ for some non-negative integers k and $l < |S||A|$, then we have:

$$
\begin{aligned}
&\max_{\{n^i_{sa}\}} \sum_{s,a} \sum_{i=1}^{N} \sqrt{\frac{n^i_{sa}}{m_0+n^{<i}_{sa}}} \\
&\leq |S||A|m \sum_{j=0}^{k-1} \frac{1}{\sqrt{m_0+jm}} + lm \frac{1}{\sqrt{m_0+km}} \\
&= (|S||A|-l)m \sum_{j=0}^{k-1} \frac{1}{\sqrt{m_0+jm}} + lm \sum_{j=0}^{k} \frac{1}{\sqrt{m_0+jm}} \\
&= \frac{|S||A|m}{\sqrt{m_0}} + (|S||A|-l)m \sum_{j=1}^{k-1} \frac{1}{\sqrt{m_0+jm}} \\
&\quad + lm \sum_{j=1}^{k} \frac{1}{\sqrt{m_0+jm}} \\
&\leq \frac{|S||A|m}{\sqrt{m_0}} + (|S||A|-l)m \int_0^{k-1} \frac{dx}{\sqrt{m_0+xm}} \\
&\quad + lm \int_0^{k} \frac{dx}{\sqrt{m_0+mx}} \\
&= \frac{|S||A|m}{\sqrt{m_0}} + 2(|S||A|-l)\sqrt{m_0+xm}\Big|_0^{k-1} \\
&\quad + 2l\sqrt{m_0+mx}\Big|_0^{k} \\
&\leq \frac{|S||A|m}{\sqrt{m_0}} + 2(|S||A|-l)\sqrt{(k-1)m} + 2l\sqrt{km} \\
&\leq \frac{|S||A|m}{\sqrt{m_0}} + 2|S||A|\sqrt{km} \\
&\leq \frac{|S||A|m}{\sqrt{m_0}} + 2\sqrt{|S||A|Nm}
\end{aligned}
$$

Putting all together, we conclude that with probability at least $1-\delta-\delta'$:

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{(s,a)\sim\rho_i}[||\hat{T}^i_{sa} - T_{sa}||_1] \\
&\leq \frac{4|S||A|}{N} + \frac{\log(2)|S|^2|A|}{Nm} + \frac{|S||A|\log(|S||A|N/\delta')}{Nm} \\
&\quad + 2\sqrt{\frac{2\log(2)|S|^2|A| + 2|S||A|\log(|S||A|N/\delta')}{Nm}} + 2\sqrt{\frac{2\log(1/\delta)}{Nm}}
\end{aligned}
$$

This is an interesting result in itself: it shows that using the empirical estimator of T at each iteration based on observed samples so far is a no-regret algorithm under this L_1 distance penalty.

Combining with the result from Lemma 4.1, this implies that for any $\epsilon > 0$, we can choose $m = 1$, $N = \tilde{O}(\frac{C^2_{\text{rng}}[|S|^2|A| + |S||A|\log(1/\delta') + \log(1/\delta)]}{\epsilon^2(1-\gamma)^4})$ to ensure that with probability at least $1-\delta-\delta'$, for any policy π':

$$
J_\mu(\hat{\pi}) \leq J_\mu(\overline{\pi}) \leq J_\mu(\pi') + \overline{\epsilon}^{\pi'}_{\text{oc}} + O(c^{\pi'}_\nu \epsilon)
$$

Thus if we solve each optimal control problem with high enough accuracy, and we have access to a good state-action exploration distribution, we can obtain an ϵ-optimal policy with high probability with sample complexity that is $O(\frac{C^2_{\text{rng}}|S|^2|A|}{\epsilon^2(1-\gamma)^4})$ (ignoring log factors). This is an improvement over other model-based RL methods that have been analyzed in this particular scenario, such as $R_{\max}$, which has sample complexity of $O(\frac{C^3_{\text{rng}}|S|^2|A|}{\epsilon^3(1-\gamma)^6})$ (Strehl et al., 2009), and a recent improved version of $R_{\max}$ which has sample complexity of $O(\frac{C^2_{\text{rng}}|S||A|}{\epsilon^2(1-\gamma)^6})$ (Szita & Szepesvári, 2010).

1.5.2. Finite MDP with Kernel SVM Model

Consider the true model to be an arbitrary finite MDP, and the set of models $\mathcal{T}$ be a set of multiclass SVM in a Reproducing Kernel Hilbert Space (RKHS) induced by some kernel k. For any state-action pair s, a, and hypothesis h in the RKHS, the associated transition model $\hat{T}^h_{sa}$ puts probability 1 on next state $s' = \text{argmax}_{s''} h(f^{s''}_{sa})$ for $f^{s'}_{sa}$ the feature vector associated with transition (s,a,s') (e.g. in a grid world domain, this might encode the relative location of s' with respect to s, direction in which a is moving the robot, and configuration of nearby obstacles or type of terrain we're on). Without loss of generality, we assume the kernel k inducing the RKHS has RKHS norm $||k(\cdot, f^{s'}_{sa})|| \leq 1$ for any transition (s,a,s') (we can scale any bounded kernel over the feature space to satisfy this), and we restrict $\mathcal{T}$ to only functions h with bounded RKHS norm $||h|| \leq K$. In the case of a linear SVM, this corresponds to assuming that the features are scaled so that $||f^{s'}_{sa}||_2 \leq 1$ and we restrict ourselves to weight vector w, such that $||w||_2 \leq K$.

To optimize the model, we consider proceeding by doing online learning on the following multiclass hinge loss functional L. Given any observed transition (s,a,s') in our dataset and SVM h, we define the loss as:

$$
\ell(h,s,a,s') = \max[0, 1 - h(f^{s'}_{sa}) + \max_{s'' \neq s'} h(f^{s''}_{sa})]
$$

We note that the loss $\ell(h,s,a,s')$ upper bounds the 0-1 classification loss $\ell_{0-1}(h,s,a,s')$ (as defined in lemma 1.3)

We will now seek to bound $\overline{\epsilon}^{\text{cls}}_{\text{prd}}$ with high probability as a function of the regret and minimum loss in the class on the sampled training data. Let $(s_{ij}, a_{ij}, s'_{ij})$ denote the j^{th} sample transition at iteration i (i.e. sampled from $\rho_i = \frac{1}{2}D_{\mu,\pi_i} + \frac{1}{2}\nu$). For $i \in \{1, 2\ldots, N\}$ and $j \in \{1, 2, \ldots, m\}$, define the random variables $Y_{(i-1)m+j} = \mathbb{E}_{(s,a)\sim\rho_i, s'\sim T_{sa}}[\ell_{0-1}(h^i, s, a, s')] - \ell_{0-1}(h^i, s_{ij}, a_{ij}, s'_{ij})$. Then $\mathbb{E}[Y_{(i-1)m+j}|Y_1, Y_2, \ldots, Y_{(i-1)m+j-1}] = 0$ and thus the random variables $X_k = \sum_{l=1}^{k} Y_l$ for $k \in \{1, 2, \ldots, Nm\}$ form a martingale.

Since $\ell_{0-1}(h,s,a,s') \in [0,1]$ for all h, s, a, s' then $|Y_l| \leq 1$ with probability 1. By Azuma-Hoeffding's inequality, we obtain that $\frac{X_{Nm}}{Nm} \leq \sqrt{\frac{2\log(1/\delta)}{Nm}}$ with probability at least $1-\delta$. Thus, using lemma 1.3, we have that with probability at least $1-\delta$:

$$
\begin{aligned}
&\overline{\epsilon}^{\text{L1}}_{\text{prd}} \\
&= 2\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{(s,a)\sim\rho_i, s'\sim T_{sa}}[\ell_{0-1}(h^i, s, a, s')] \\
&\leq 2[\frac{1}{Nm}\sum_{i=1}^{N}\sum_{j=1}^{m} \ell_{0-1}(h^i, s_{ij}, a_{ij}, s'_{ij}) + \sqrt{\frac{2\log(1/\delta)}{Nm}}] \\
&\leq 2[\frac{1}{Nm}\sum_{i=1}^{N}\sum_{j=1}^{m} \ell(h^i, s_{ij}, a_{ij}, s'_{ij}) + \sqrt{\frac{2\log(1/\delta)}{Nm}}]
\end{aligned}
$$

Now with these samples, the online algorithm is run on the sequence of loss functionals $L_i(h) = \frac{1}{m}\sum_{j=1}^{m} \ell(h, s_{ij}, a_{ij}, s'_{ij})$. Because the L_i are all convex

in h, for any (s, a, s'), then an online algorithm such as gradient descent, or follow-the-regularized-leader is no-regret. In particular, suppose we run the projected subgradient descent algorithm from Zinkevich (2003). Because for any h, h' in the RKHS, $||h - h'|| \leq 2K$ and for any h, s, a, s', the norm of the subgradient $||\nabla L|| = ||k(\cdot, f_{sa}^{s'}) - k(\cdot, f_{sa}^{s^*})|| \leq 2$ (for $s^* = \text{argmax}_{s'' \neq s'} h(f_{sa}^{s''})$), then using learning rate $\frac{K}{\sqrt{n}}$ at iteration n we can guarantee that $\frac{1}{N}\sum_{i=1}^{N} L_i(h^i) \leq \min_h \frac{1}{N}\sum_{i=1}^{N} L_i(h) + \frac{6K}{\sqrt{N}}$ from the result in Zinkevich (2003). Let $\hat{\epsilon}_{\text{mdl}}^{\text{cls}} = \min_h \frac{1}{Nm}\sum_{i=1}^{N}\sum_{j=1}^{m} \ell(h, s_{ij}, a_{ij}, s'_{ij})$ the predictive error of the best model in hindsight on the training set. Then combining with the previous equation, we obtain that with probability at least $1 - \delta$:

$$\bar{\epsilon}_{\text{prd}}^{\text{L1}} \leq 2[\hat{\epsilon}_{\text{mdl}}^{\text{cls}} + \frac{6K}{\sqrt{N}} + \sqrt{\frac{2\log(1/\delta)}{Nm}}]$$

Combining with the result from Lemma 4.1, this implies that for any $\epsilon > 0$, we can choose $m = 1$, $N = O(\frac{C_{\text{rng}}^2(K^2 + \log(1/\delta))}{\epsilon^2(1-\gamma)^4})$ to ensure that with probability at least $1 - \delta$, for any policy π':

$$J_\mu(\hat{\pi}) \leq J_\mu(\overline{\pi}) \leq J_\mu(\pi') + \bar{\epsilon}_{\text{oc}}^{\pi'} + 2c_\nu^{\pi'} H \hat{\epsilon}_{\text{mdl}}^{\text{cls}} + O(c_\nu^{\pi'}\epsilon)$$

Thus if we solve each optimal control problem with high enough accuracy, there exist a SVM model in the RKHS that achieves low enough loss on the training set, and we have access to a good state-action exploration distribution, we can obtain a ϵ-optimal policy with high probability with sample complexity that is $O(\frac{C_{\text{rng}}^2 K^2}{\epsilon^2(1-\gamma)^4})$ (ignoring log factors). Note that this has no dependency on $|S|$ and $|A|$, only on the complexity of the class of models (i.e. K), which could be constant as $|S|$, $|A|$ increases.

References

Strehl, A. L., Li, L., and Littman, M. L. Reinforcement learning in finite MDPs: PAC analysis. *JMLR*, 2009.

Szita, I. and Szepesvári, C. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, 2010.

Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2003.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.