IRL with Positive and Negative Labels

James MacGlashan

1 Problem Statement

In this document we propose a variant of the inverse reinforcement learning problem. The learning problem is defined as being given a dataset of labeled trajectories and outputting a reward function, where each trajectory is a sequence of state-action pairs and each trajectory label is a positive ($+1$), or negative (-1) label that indicates whether the associated trajectory is a good example of some desired behavior, or a bad example of some desired behavior. The goal is to recover a reward function that motivates good behavior. In particular, we adopt the probabilistic maximum likelihood formulation, in which we seek a reward function that maximizes the likelihood of the data.

2 Probability Model

To find the reward function that maximizes the likelihood of the data, we first define probability model of the system that is represented in Figure 1. The motivation for this model is that we consider the the final label (L) that a user gives a trajectory of size N to be some random function of what they thought about each of the action selections (A) exhibited in the trajectory. However these step-wise evaluations (X) are unobserved in the data.

We model the probability that an action is evaluated as good or not as proportional to its selection probability according a softmax policy computed for the reward function with parameters $\boldsymbol{\theta}$. Specifically:

$$\Pr(x_i = +1 | s, a, \theta) = \pi(s, a|\boldsymbol{\theta}) \tag{1}$$

$$\Pr(x_i = -1 | s, a, \theta) = 1 - \pi(s, a|\boldsymbol{\theta}), \tag{2}$$

where $\pi(s, a|\boldsymbol{\theta})$ is the softmax policy over Q-values computed for the reward function parameterized by $\boldsymbol{\theta}$:

$$\pi(s, a|\boldsymbol{\theta}) = \frac{e^{\beta Q(s,a|\boldsymbol{\theta})}}{\sum_{a'} e^{\beta Q(s,a'|\boldsymbol{\theta})}}, \tag{3}$$

β is a selectable parameter, and $Q(s, a|\boldsymbol{\theta})$ is the Q-value computed for the reward function parameterized by $\boldsymbol{\theta}$.

For the probability distribution of L, given the sequence of N step-wise labels, we would like a distribution that has the property that as more step-wise

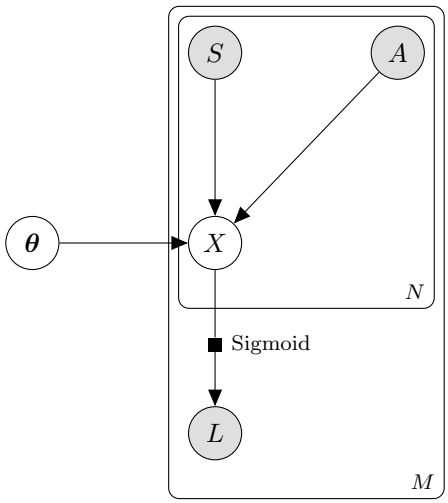

Figure 1: Plate Diagram of our Probability Model

labels are positive, the probability of a positive trajectory label increases (and vice versa). Although there are many possible distributions that satisfy this property, for concreteness, we choose the sigmoid function. That is,

$$\Pr(L = +1|X_1, ..., X_n) = \frac{1}{1 + e^{-\phi \sum_i^N X_i}} \tag{4}$$

$$\Pr(L = -1|X_1, ..., X_n) = 1 - \Pr(L = +1|X_1, ..., X_n), \tag{5}$$

where ϕ is a selectable parameter that tunes how quickly of a majority of step-wise labels increases/decreases the probability of trajectory assignment. For example, when $\phi = 0$, trajectory labels are assigned uniformly randomly independently of step-wise labels. As $\phi \to \infty$, the sigmoid converges to a step function in which a trajectory containing even one more positive step-wise label than negative step-wise labels will deterministically cause a positive trajectory label (and vice versa).

3 Finding the Parameters that Maximum the Likelihood

Recall that our goal is to find the maximum likelihood reward function parameters given a dataset with M trajectories (sequences of S, A pairs) and trajectory labels (L), but *not* step-wise labels (X). Under our model, the likelihood of such

a dataset D of size M for a given set of reward function parameters θ is

$$\begin{aligned}
\mathcal{L}(D|\boldsymbol{\theta}) &= \prod_{i}^{M} \Pr(l_i, s_{i,1}, a_{i,1} ..., s_{i,N}, a_{i,N}|\boldsymbol{\theta}) \\
&= \prod_{i}^{M} \sum_{\mathbf{x}\in\{-1,1\}^N} \Pr(l_i, s_{i,1}, a_{i,1}, x_1, ..., s_{i,N}, a_{i,N}, x_N|\boldsymbol{\theta}) \\
&= \prod_{i}^{M} \sum_{\mathbf{x}\in\{-1,1\}^N} \Pr(l_i|x_1, ..., x_n) \Pr(x_1, ..., x_N|s_{i,1}, a_{i,1}, ..., s_{i,N}, a_{i,N}, \boldsymbol{\theta}) \\
&= \prod_{i}^{M} \sum_{\mathbf{x}\in\{-1,1\}^N} \Pr(l_i|x_1, ..., x_n) \prod_{j}^{N} \Pr(x_j|s_{i,j}, a_{i,j}, \boldsymbol{\theta}) \qquad (6)
\end{aligned}$$

We will estimate the maximum likelihood parameters by using gradient ascent to (locally) maximize the log likelihood function, where the log likelihood is:

$$\begin{aligned}
\log \mathcal{L}(D|\boldsymbol{\theta}) &= \sum_{i}^{M} \log \sum_{\mathbf{x}\in\{-1,1\}^N} \Pr(l_i|x_1, ..., x_n) \prod_{j}^{N} \Pr(x_j|s_{i,j}, a_{i,j}, \boldsymbol{\theta}) \\
&= \sum_{i}^{M} q_i + \log \left[\sum_{\mathbf{x}\in\{-1,1\}^N} \exp\left(-q_i + \log\left(\Pr(l_i|x_1, ..., x_n) \prod_{j}^{N} \Pr(x_j|s_{i,j}, a_{i,j}, \boldsymbol{\theta}) \right) \right) \right] \\
&= \sum_{i}^{M} q_i + \log \left[\sum_{\mathbf{x}\in\{-1,1\}^N} \exp\left(-q_i + \log\left(\Pr(l_i|x_1, ..., x_n)\right) + \sum_{j}^{N} \log\left(\Pr(x_j|s_{i,j}, a_{i,j}, \boldsymbol{\theta})\right) \right) \right],
\end{aligned}$$

where

$$q_i = \max_{\mathbf{x}\in\{-1,1\}^N} \log\left(\Pr(l_i|x_1, ..., x_n)\right) + \sum_{j}^{N} \log\left(\Pr(x_j|s_{i,j}, a_{i,j}, \boldsymbol{\theta})\right).$$

To simplify the expression of this function, and make the gradient more clear, we will break the log likelihood function up into different sub-functions,

which are all implicitly defined with respect to a $\boldsymbol{\theta}$. Let

$$P_{xij} = \begin{cases} \pi(s_{i,j}, a_{i,j}|\boldsymbol{\theta}) & \text{if } x_{i,j} = 1 \\ 1 - \pi(s_{i,j}, a_{i,j}|\boldsymbol{\theta}) & \text{if } x_{i,j} = -1 \end{cases} \tag{7}$$

$$net_x = \phi \sum_j^N x_j \tag{8}$$

$$S(x) = \frac{1}{1 + e^{-x}} \tag{9}$$

$$P_{lix} = \begin{cases} S(net_{xi}) & \text{if } l = 1 \\ 1 - S(net_{xi}) & \text{if } l = -1 \end{cases} \tag{10}$$

$$g_{xi} = \log P_{lix} + \sum_j^N \log P_{xij} \tag{11}$$

$$m_i = \underset{\mathbf{x} \in \{-1,1\}^N}{\arg\max}\, g_{xi} \tag{12}$$

$$f_{xi} = g_{xi} - g_{m_i i}. \tag{13}$$

Then, using these terms, we can rewrite the log likelihood as

$$\log \mathcal{L}(D|\boldsymbol{\theta}) = g_{m_i i} + \log \sum_{\mathbf{x} \in \{-1,1\}^N} e^{f_{xi}}. \tag{14}$$

The partial derivative for any single parameter θ of $\boldsymbol{\theta}$ of the log likelihood under this expression is

$$\frac{\partial}{\partial \theta} \log \mathcal{L}(D|\boldsymbol{\theta}) = \frac{\partial}{\partial \theta} g_{m_i i} + \frac{\sum_{\mathbf{x} \in \{-1,1\}^N} \left(e^{f_{xi}}\right) \frac{\partial}{\partial \theta} f_{xi}}{\sum_{\mathbf{x} \in \{-1,1\}^N} e^{f_{xi}}}. \tag{15}$$

We will compute the partial partial derivatives for each of these terms by ex-

amining the partial derivatives for each of our sub functions, bottom up.

$$\frac{\partial}{\partial\theta} net_x = 0 \tag{16}$$

$$\begin{aligned}\frac{\partial}{\partial\theta} S(net_x) &= S(net_x)(1 - S(net_x))\frac{\partial}{\partial\theta} net_x \\ &= S(net_x)(1 - S(net_x))0 \\ &= 0\end{aligned} \tag{17}$$

$$\begin{aligned}\frac{\partial}{\partial\theta} P_{lix} &= \begin{cases} \frac{\partial}{\partial\theta} S(net_x) & \text{if } l = 1 \\ -\frac{\partial}{\partial\theta} S(net_x) & \text{if } l = -1 \end{cases} \\ &= 0\end{aligned} \tag{18}$$

$$\frac{\partial}{\partial\theta} P_{xij} = \begin{cases} \frac{\partial}{\partial\theta}\pi(s_{i,j}, a_{i,j}|\boldsymbol{\theta}) & \text{if } x_{i,j} = 1 \\ -\frac{\partial}{\partial\theta}\pi(s_{i,j}, a_{i,j}|\boldsymbol{\theta}) & \text{if } x_{i,j} = -1 \end{cases} \tag{19}$$

$$\begin{aligned}\frac{\partial}{\partial\theta} g_{xi} &= \frac{\frac{\partial}{\partial\theta} P_{lix}}{P_{lix}} + \sum_j^N \frac{\frac{\partial}{\partial\theta} P_{xij}}{P_{xij}} \\ &= \sum_j^N \frac{\frac{\partial}{\partial\theta} P_{xij}}{P_{xij}}\end{aligned} \tag{20}$$

$$\frac{\partial}{\partial\theta} f_{xi} = \frac{\partial}{\partial\theta} g_{xi} - \frac{\partial}{\partial\theta} g_{m_i i}. \tag{21}$$

It is interesting to note that the partial derivative of the probability distribution for the L function goes to zero. This result is due to the fact that once some set of X values have been selected, the likelihood does not depend on $\boldsymbol{\theta}$, and in the likelihood function, X values are bound during the marginalization process in which we sum over them. However, it should be noted that just because the partial derivative for L is zero, that does not mean that its probability distribution plays no role in the full partial derivative of the log likelihood. You will find the probability of L given the Xs non-differentiated value appearing in the exponentiation of our f_{xi} terms.

Finally, in this document, I did not provide the partial derivative of the policy, which in turn depends on the partial derivative of the Q-function. However, these terms have been explained in existing IRL work that we can reuse.

4 Importance Sampling

Recall likelihood.

$$\mathcal{L}(D \mid \theta) = \prod_i^M \sum_{\mathbf{x}\in\{-1,1\}^N} \Pr(l_i|x_1, ..., x_n) \prod_j^N \Pr(x_j|s_{i,j}, a_{i,j}, \boldsymbol{\theta}) \tag{22}$$

Is also equal to...

$$\mathcal{L}(D \mid \theta) = \prod_i^M \sum_{\mathbf{x}\in\{-1,1\}^N} \Pr(l_i|x_1,...,x_n) \prod_j^N \Pr(x_j|s_{i,j},a_{i,j},\boldsymbol{\theta}) \frac{\zeta(\boldsymbol{x})}{\zeta(\boldsymbol{x})} \tag{23}$$

Now we Monte-Carlo sample according to q

$$\hat{\mathcal{L}}(D \mid \theta) = \prod_i^M \frac{1}{C} \sum_k^C \frac{\Pr(l_i|x_1,...,x_n) \prod_j^N \Pr(x_j|s_{i,j},a_{i,j},\boldsymbol{\theta})}{\zeta(\boldsymbol{x})} \tag{24}$$

Now we do log likelihood

$$\begin{aligned}
\log \hat{\mathcal{L}}(D|\boldsymbol{\theta}) &= \sum_i^M \log \frac{1}{C} \sum_k^C \frac{\Pr(l_i|x_1,...,x_n) \prod_j^N \Pr(x_j|s_{i,j},a_{i,j},\boldsymbol{\theta})}{\zeta(\boldsymbol{x})} \\
&= \sum_i^M \log \frac{1}{C} + q_i + \log \left[\sum_k^C \exp \left(-q_i + \log \left(\frac{\Pr(l_i|x_1,...,x_n) \prod_j^N \Pr(x_j|s_{i,j},a_{i,j},\boldsymbol{\theta})}{\zeta(\boldsymbol{x})} \right) \right) \right] \\
&= \sum_i^M \log \frac{1}{C} + q_i + \log \left[\sum_k^C \exp \left(-q_i - \log \zeta(\boldsymbol{x}) + \log \left(\Pr(l_i|x_1,...,x_n) \right) + \sum_j^N \log \left(\Pr(x_j|s_{i,j},a_{i,j},\boldsymbol{\theta}) \right.\right.\right.
\end{aligned}$$

where,

$$q_i = \max_{\text{samples}} -\log \zeta(\boldsymbol{x}) + \log \left(\Pr(l_i|x_1,...,x_n) \right) + \sum_j^N \log \left(\Pr(x_j|s_{i,j},a_{i,j},\boldsymbol{\theta}) \right).$$