ReNeg (Regression with Negative Examples) For Immitation Learning with Autonomous Vehicles

Jacob Beck
Brown University RLab, Self-Driving Car (SDC) Lab
In collaboration with Zoe Papakipos
Advised by Professor Michael Littman

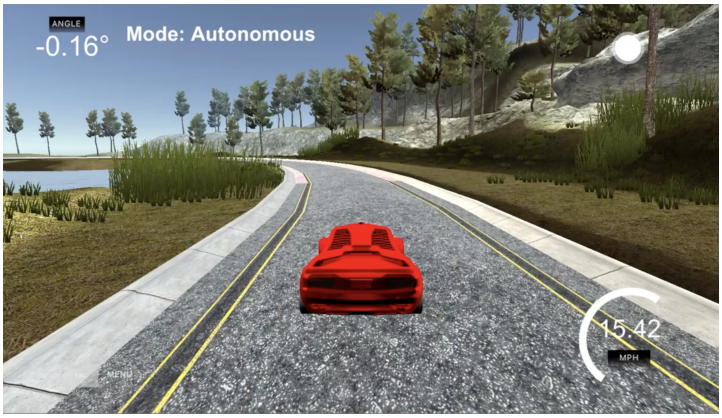

Fig. 1. Self-Driving Car Simulation

I. ABSTRACT

Since reinforcement learning (RL) involves making mistakes, it is only viable for autonomous vehicle (AV) control in a simulation. Therefore, machine learning for autonomous vehicle control in the real world focuses either on a subset of the problem (i.e. computer vision) or behavioral cloning (i.e. supervised learning from a correct demonstration). We would like to propose a third alternative that will allow for better and faster learning of any subset of the entire problem: learning from demonstration that covers the full range of positive to negative behavior. We expect to find that the additional information provided by incorrect behavior (and by gradations in how good or bad a behavior is) will enable the autonomous agent to learn the correct continuous control output (policy) more quickly. In our work we assume our learned policy cannot be tested as it learns, but we do have access to demonstrated actions from an expert, which intentionally include both good and bad examples, in addition to feedback on how good or bad those actions are in expectation. Existing work in the field is generally not capable of solving this problem for one of three reasons:

1) The work requires the agent to explore and make mistakes (a la DAgger [1], AggreVaTeD [2], and Deeply AggreVaTeD [3]) to make use of the expert feedback.
2) The work is set in the off-policy RL framework (a la Off-Policy Actor-Critic [4], Q-Learning [5], Retrace [6]). These approaches assume the behavior policy used to explore has a non-zero probability of choosing each possible action in any given state, which will not work in a potentially dangerous AV setting, and also these approaches tend to modify the objective function into an expectation over the states visited by the behavior policy, which will not be an accurate assumption for the agent's learned policy at runtime.

3) The work focuses on discrete action control output instead of continuous control and also requires entropy in the objective function which could be dangerous in the context of AV (a la Normalized Actor Critic, or NAC) [7].

The main contributions of this paper are the following:

- To propose a problem statement and motivate its need: learning deterministic and continuous control from good and bad demonstration, coupled with expert feedback.
- To empirically validate various solutions in the context of autonomous vehicle control on limited data, using transfer learning.
- Where appropriate, to draw connections to, and inspiration from, the RL framework.

Moreover, we find that the empirically best solution in this context is simple and has strong connections to stochastic policy gradient approaches, and so our solution is non-brittle with respect to the hyper parameter tuning of more complicated neural networks and could be easily extended to RL with no changes to the update rule.

II. BACKGROUND

Learning from demonstration has been studied in the supervised learning framework and the Markov Decision Process (MDP) framework. In the former, it is generally known as imitation learning (or behavioral cloning in the AV setting), and it the latter, it is generally known as apprenticeship learning. The supervised learning in this area generally amounts to a least squares regression that maps from an input state to action, with some research addressing the the fact that our runtime distribution will differ from our training distribution. (We will discuss this later.) The MDP research here has largely focused on apprenticeship learning via inverse reinforcement learning (AIRL) [8], in which a reward function is estimated given the demonstration of an expert. This reward function is estimated via an iterative approach where a reward function is assigned that minimizes the differences between the value induced by the current policy and the expert policy, and then the current policy is improved to maximize the current reward function. However, in this framework, the reward function is restricted to the class of linear combinations of the discrete features of the state, and the framework only allows for positive examples. In addition, there has been work on inverse reinforcement learning from failure (IRLF) [9], which allows for a sequence of positive or a sequence of negative examples, however it still does not allow for any of the following: an arbitrary reward function on continuous inputs, gradation in how good or bad the demonstration is, labels for atomic actions as opposed to a trajectory or sequence of actions.

Perhaps the closest framework that is a candidate for our work is off-policy gradients for reinforcement learning [10]. We will note in this paper when we use ideas inspired by this framework, however, it still does not fit for the following reasons: 1) We would like to recover supervised behavioral cloning if only "positive" data happens to be found in our labels. This is fairly critical for certain driving scenarios where only the correct action can be demonstrated. (For example, during a very tight and well-timed dangerous manoeuvre). However, most of these off-policy methods cannot recover behavioral cloning since they cannot operate with deterministic policies: they require their behavior policy, which is the policy used for exploration, to have a non-zero chance of executing any possible trajectory that the learned policy can execute, or their convergence requires that any given action in any given state has a non-zero probability so that each (state, action) pair can be visited an infinite number of times in the limit [11] [7]. 2) To this end, the sign of our labelled feedback is very important, but the distinction between good and bad examples does not exist in the objective function

for reinforcement learning. 3) Ideally, we would have a parameter that we can adjust to select how much we would like to clone positive examples as opposed to avoid negative examples, 4) Although we will belabor this point momentarily for both RL and supervised learning, off-policy RL approaches tend to modify the objective function into an expectation over the states visited by the behavior policy, which will not be an accurate assumption for the agent's learned policy at runtime.

A method for behavioral cloning with end-to-end behavioral cloning was put forth by NVIDIA [12], however, to the best of our knowledge, no research so far has focused on using expert-labelled feedback in the context of autonomous vehicle control. (Nor even regression with negative examples in general, as mentioned earlier.) Yet, we believe that this is a major oversight for self-driving cars: given that many research groups are investing exorbitant amounts of time into collecting driving data, if it is truly the case that our hypothesis is correct and that labelling feedback will help the car to learn, such gains can be gotten simply by having an expert labeller sit in the car alongside the driver. For no additionally real world time, advantages can be gained simply by having a "backseat driver".

Our problem statement is somewhere in between behavioral cloning and RL: we focus on a general approach that is capable of mapping continuous sensor input to an arbitrary (differentiable) continuous policy output and capable of using feedback for singular predetermined actions. We believe that this information is easily collected in our (and many) contexts and constitutes a stronger learning signal. Moreover, we take our feedback to include much more information than just a reward: we take our label to directly measure how "good" an action is relative to other possible actions. We use this approach instead of labelling rewards for actions since we found it easy to label data in this way with minimal practice, and it is far more information-rich.

Because our policies, as in behavioral cloning, do not get to "explore" by choosing actions, we are learning "off-policy". There are two issues here: first, we must find a loss function that leads to a useful policy; second, the runtime distribution will be different from the training distribution. The first issue will be the focus of the paper and will be discussed at length. We will enumerate desired properties of the loss function and explain which we can attain with each loss function we use for an experiment. However, we additionally believe our ReNeg approach will likely help mitigate the second prevalent in supervised behavioral cloning: namely, that the expert distribution often differs from the runtime distribution induced by the learned policy (that is an approximation of the expert). Typically, an expert will encounter a very different set of states than the agent that we train, since the agent will not be as good as the expert. Once the trained agent is in these states that it has not seen, the issue is compounded. Thus the assumption that our training data is independent and identically (i.i.d.) distributed from the actual distribution goes out the window. One way to mitigate this issue would be to have the expert collect data in "bad" states that it would generally not reach, as well as states that its distribution would generally induce. That is, in our setting, have the expert driver label states from all over the road. However, this is impractical and would require the driver to record "bad" data as they are on their way to one of these states that their policy would not generally see. However, since we can collect positive and negative examples, this is not an issue for ReNeg: we can simply label these states as "bad" and record them for training along with our "good" data. Thus, ReNeg frees up our expert to easily expose the training data to bad states that are off of their own policy by explicitly labelling these states as "bad".

Two things should be noted here: 1) In the off-policy policy gradient RL framework, this issue seems to be primarily circumvented by changing the objective function from an expectation of the learned policy's value function over the learned policy state-visitation distribution to an expectation of the learned policy's value function over the behavior (exploratory) state-visitation distribution [4]. 2)

There is a well known solution to this problem in the context of supervised learning: Stephane Ross el al. present a solution known as DAgger, that allows the agent to explore and then uses the expert to label the new dataset, then training on all of the combined data [1]. Such an approach has even been improved upon both to address RL by incorporating experts that can label Q values with the AgraVaTeD algorithm [2], and to and to address deep neural networks by finding a policy gradient with the Deeply AgraVaTeD algorithm [3].

These approaches, however, will not work as the only solution for us, since the entire goal was to avoid allowing the agent to make mistakes while driving. Thus, we need to expose it to all states from the start, as much as possible, without allowing the agent to make decisions. Finding a representation and loss function to take care of deterministic off-policy learning for continuous control during the training phase is the goal of this paper. It might be suggested that, once we have trained our agent using ReNeg, we could switch to a DAgger inspired method such as Deeply AgraVaTeD could be used for supplemental fine-tuning on the correct runtime distribution, and perhaps this is true, but it should also be noted that, if our policy were actually this good, it would also land us much more squarely in the on-policy reinforcement learning framework, at which point we would have to weight the benefits of a stochastic policy gradient approach with an extremely low variance (to keep the trajectories safe), using a deterministic policy-gradient, or continuing to view the feedback as $Q*$ values and use an approach such as Deeply AgraVaTeD. In the AV setting, the primary disadvantage of Deeply AgraVaTeD would be that it requires mixing an expert policy with the current learned policy, and so a human would have to be attentively steering as well, in case the action is sampled from the human.

For the task of self-driving cars, we chose to learn our policy with an end-to-end optimization of a neural network to approximate a continuous control function. We believe that such network are capable of tackling a wide array of general problems, and that, When end-to-end learning works, it has the potential to better optimize all aspects of the pipeline for the task at hand. If the option is to use separate (differentiable) pre-made modules for certain sub-tasks such as computer vision and not optimize them for the given task, or to connect these modules in a way that can be updated for the given task, the latter will always perform better, since the end-to-end model always has the ability to simply not update the sub-module if it will not help training loss. We note that having a black box be entirely responsible for controlling our cars is rightly a common concern, but we will enumerate our ideology and motivation with respect to the autonomous vehicle industry, in the hope that ReNeg can still contribute to this field: 1) black box algorithms may perform better provided enough validation and additional concurrent sanity checks, 2) there may be situations in which black box algorithms are the only system capable of confidently preventing imminent danger, 3) we are hopeful that neural networks will become less of a black box either through additional visualization research, especially when the input is visually interpretable, or by design, by injecting loss into earlier neurons to intentionally create their purpose, and learning a gating mechanism over such neurons 4) we believe that ReNeg can be used for smaller sub-problems in autonomous vehicle planning if there are areas that practitioners feel comfortable controlling with a gradient-based approach.

III. METHODS

The task we will be evaluating our model on is is lane-following in a car simulator implemented in Unity. The correctness of an example will be labelled using feedback from a human. We use expert feedback as opposed to expert answers (explicit steering angles) simply because it is far easier for an expert to give correct feedback when they themselves are not driving, as opposed to needing to know the exact angle at which to turn the wheel to make a turn when you can't see how your actions affect the car. Our goal is to produce a policy network (here on abbreviated as PNet) to map states to actions.

In our case, the states will be RGB images from the driver's point of view, and the PNet will have to output a steering angle that keeps the car as close to the center of the road as possible. The crux of the issue we are tackling is finding the correct loss function to take in positive and negative examples with gradations in both. Secondarily, it is important to choose the feedback in an appropriate way as to encourage the correct behavior in the context of lane following. And finally, we will discuss architecture implementation.

A. Loss Function

If θ is the angle in the demonstrated example, $\hat{\theta}$ is the angle predicted by the PNet, and f is the feedback, then the loss function we choose should have the following 3 properties:

1) Minimizing the loss should minimize the distance between θ and $\hat{\theta}$ for positive examples. That is, for positive examples, the loss should increase with the distance between θ and $\hat{\theta}$
2) Minimizing the loss should should maximize the distance between θ and $\hat{\theta}$ for negative examples. That is, for negative examples, the loss should decrease with the distance between θ and $\hat{\theta}$
3) The degree to which minimizing the loss does 1) and 2) should be determined by the magnitude of the feedback. That is, the magnitude of the loss should increase with the magnitude of the f

These three properties together will ensure that the network avoids the worst negative examples as much as possible, while seeking the best examples. Given an input state s, the first loss function that comes to mind is what we term "scalar loss":

$$Loss_{scalar} = f * (\theta(s) - \hat{\theta}(s))^2$$

This loss function is notable for several reasons. The first reason that this loss is notable is that it looks a lot like the mean squared loss used for behavioral cloning: $Loss_{clone} = (\theta - \hat{\theta})^2$. In fact, we have set up two hyper-parameters that, when set appropriately, will recover this behavioral cloning loss. The first parameter that we introduced to our loss function was a Boolean setting called THRESHOLD, or THRESH for short. If THRESH is set to true, we simply replace f with $sign(f)$. This will eliminate gradations in positive and negative data and will provide a good metric for comparison. Additionally, we introduced the parameter α. We replace f with $max(\alpha * f, f)$. This has the effect of scaling down all of our negative feedback by α and is useful in isolating the effects of gradations in all data from gradations in just positive data. We apply α after we threshold so that with THRESH set to true and an α value of 0.0, we an can recover behavioral cloning.

The second reason our scalar loss is notable is that it closely resembles the loss that induces a stochastic policy gradient in standard continuous control reinforcement learning. In a standard RL policy network such as REINFORCE, the loss function would be $Loss = R * -log(Pr(\theta))$, which encourages the probability of the action taken by an amount proportional to an unbiased sample of future discounted reward (or return), R [13]. In continuous control, you would instead predict a mean $\hat{\theta}$ for a normal distribution and then sample you action θ from that normal distribution. Now if you replace $log(Pr(\theta))$ with the probability density function for a normal distribution, the loss you wind up with looks a lot like our scalar loss: $Loss = R * (\theta - \hat{\theta})^2$ (ignoring some constants based on the variance of the normal distribution that don't depend on θ or $\hat{\theta}$ and so only affect learning rate). Here is the math to go from the stochastic policy gradient to the scalar loss that induces it:

$$\nabla Loss = \nabla R * -log(Pr(\theta))$$

$$\nabla Loss = \nabla R * -log(\frac{e^{-\frac{(\theta-\hat{\theta})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}})$$

$$\nabla Loss = \nabla R * -(log(e^{-\frac{(\theta-\hat{\theta})^2}{2\sigma^2}}) - log(\sqrt{2\pi\sigma^2}))$$

$$\nabla Loss = \nabla R * -(log(e^{-\frac{(\theta-\hat{\theta})^2}{2\sigma^2}})$$

$$\nabla Loss = \nabla R * \frac{(\theta-\hat{\theta})^2}{2\sigma^2}$$

$$\nabla Loss \propto \nabla R * (\theta-\hat{\theta})^2$$

$$Loss \propto R * (\theta-\hat{\theta})^2$$

Although this similarity provides inspiration, it is in fact not justified in this context by reinforcement learning. Since we are "off-policy", the neural network cannot not influence the probability of seeing an example again, and this leads can lead to problems. In RL, the network could try a bad action, and then would move away from it and not revisit it. Whereas, if we have a bad example in our training set for a given state, on every epoch of training, our neural net will encounter this example and take a step away from it, thus pushing our network as far away from it as possible. In fact, even if we have a positive example for that very state, if we have more negative examples than positive examples, if we are not careful, we may wind up ignoring our positive examples completely in an effort to get away from our negative examples. This case highlights the trouble inherent in using negative examples: It is hard to know how and when to take into account the negative examples and by how much.

In the RL framework, this could be dealt with by an approximation off-policy stochastic policy gradient that scales the stochastic policy gradient by an importance sampling ratio [10] [4]. If we make a few assumptions, we may be able to use this gradient, but let me provide the assumptions and then some reasons why we did not use this gradient. First, we would have to assume that when we go to use our learned policy, we will use the predicted value as a mean for a stochastic runtime distribution, which is not something we want to do. Second, we would have to assume that we know the probability with which the expert human driver took their action. We could assume it was done deterministically, but then this violates the condition necessary in exploration that the behavior policy not have 0 probability of choosing an action. Third, the updates will not work well with ReNeg's notion of positive and neagative examples: for data points with positive feedback with large errors has an insignificant update, as well as data points with small errors and negative feedback. This will, at very least, require an exorbitant number of gradient updates. The following is a concrete example: Let's say that our policy "would" choose actions by sampling from a Gaussian distribution centered at the predicted action, with a variance of $2^{-1/2}$. Let the probability density function of such a normal distribution for a point x be written as $G(\mu, x)$, and let $SG(x)$ represent a "stop gradient" operation that turns x into a constant instead of a function. Note that proportional loss functions will induce a proportional gradient based update, and so the difference can be accounted for entirely by the learning rate. Then, a loss function we could use to induce the correct gradient would be:

$$Loss = SG(\frac{P(\theta|LearningPolicy)}{P(\theta|BehaviorPolicy)}) * R * (\theta-\hat{\theta})^2$$

$$Loss = SG(\frac{G(\hat{\theta},\theta)}{1}) * R * (\theta-\hat{\theta})^2$$

$$Loss = SG(\frac{e^{-\frac{(\theta-\hat{\theta})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}) * R * (\theta-\hat{\theta})^2$$

$$Loss = SG(\frac{e^{-\frac{(\theta-\hat{\theta})^2}{2(2^{-1/2})^2}}}{\sqrt{2\pi(2^{-1/2})^2}}) * R * (\theta - \hat{\theta})^2$$

$$Loss = SG(\frac{e^{-(\theta-\hat{\theta})^2}}{\sqrt{\pi}}) * R * (\theta - \hat{\theta})^2$$

$$Loss \propto SG(\frac{1}{e^{(\theta-\hat{\theta})^2}}) * R * (\theta - \hat{\theta})^2$$

However, here are the additional and primary concern with such a loss function: 1) If we consider a simple case where we are trying to learn the continuous angle action 0 from one data point that labels the action 0 with a feedback of f, and we have exactly 1 parameter, which is the current $\hat{\theta}$, then we can see that our update to theta will be dwarfed by the denominator in the importance sampling ratio. Given a learning rate α, the update would be:

$$\hat{\theta} := \hat{\theta} - \alpha * \frac{1}{e^{(\theta-\hat{\theta})^2}} * R\nabla(\theta - \hat{\theta})^2$$

$$\hat{\theta} := \hat{\theta} - \alpha * \frac{1}{e^{\hat{\theta}^2}} * R\nabla\hat{\theta}^2$$

$$\hat{\theta} := \hat{\theta} - \alpha * \frac{1}{e^{\hat{\theta}^2}} * R * 2 * \hat{\theta}$$

Removing constants that would be the same for every piece of data, this makes the difference for the update proportional to:

$$\propto -\frac{f * \hat{\theta}}{e^{\hat{\theta}^2}}$$

Once $\hat{\theta}$ gets large, we can see that this loss function actually decreases the amount we update $\hat{\theta}$ as our estimate gets worse and worse! We will discuss soon how to introduce a new desired property to ensure we do not consider such loss functions.

We will now consider negative examples back in the ReNeg framework. For example, in Figure 2, if we perform a regression on positive and negative examples, with more negative examples in one state, we may wind up in a case where our loss is minimized by an output of positive or negative infinity, and our regression is "overwhelmed" by the negative examples.

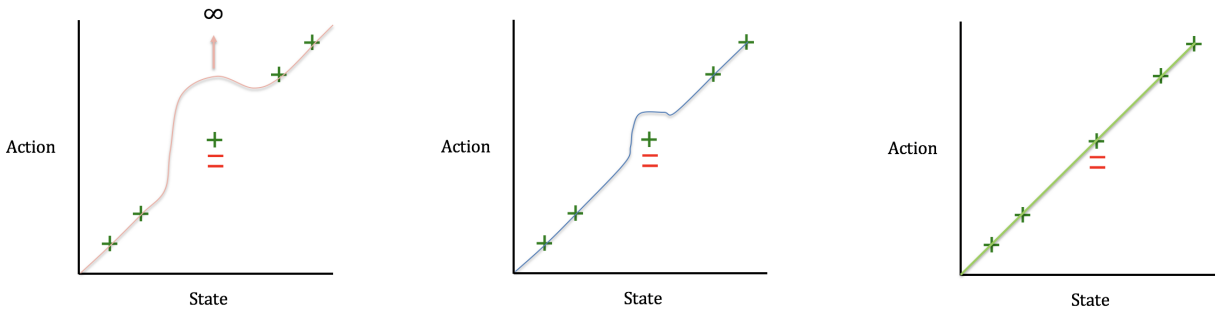

Fig. 2. Potential Outcomes of Regression

It is nice that in our current loss function, our negative examples make the loss function grow no faster with the distance between θ and $\hat{\theta}$ than do the positive examples, but we can do better than that. It would be ideal if the rate of growth of the loss slowed slowed for negative examples as the distance increased. In an effort to avoid this, we introduce a 4th desired property:

4) The magnitude of the loss should grow at least linearly with the distance between θ and $\hat{\theta}$ for positive examples. The magnitude of the loss should grow less than linearly with the distance between θ and $\hat{\theta}$ for negative examples. This enforces that, positive examples that are far away should matter at least proportionally as much as close positive examples, and negative examples that are far away should matter less than negative examples close by

This led us to our second "exponential" loss:

$$Loss_{exp} = |\theta(s) - \hat{\theta}(s)|^{2f}$$

Using this loss, negative examples will have infinite loss at distance 0, and then drop off exponentially with distance. We hope that this will create regressions more akin to the second image in Figure 2. In this image, adding more negative points will still nudge the regression away more and more, but one positive point not too close by should be enough to prevent it from converging to positive or negative infinity. It should be noted, that the loss in a particular state still could only have negative examples, especially in a continuous state-space like ours where states are never truly repeated, however the reduction in loss caused by converging towards infinity would be so small that it should not happen simply due to the continuity with nearby states enforced by the structure of the network. In addition, one concern with this loss could be that for positive fractional differences, and negative non-fractional differences, the desired property 3) of loss functions no longer holds. That is, our positive loss will not grow with f it the difference being exponentiated is a fraction. And for negative exponents, the loss will only grow if the fractional denominator is a fraction that shrinks as it is raised to increasing powers of f. However, we hope that for negative examples, distances that are more than 1 unit away will not matter much (since, as discussed later, 1 unit is half the entire range of our output). And, for both positive and negative examples, I will later propose a solution to patch this loss function for future work.

Our final loss function is intended to produce regressions more like the final image in Figure 1, but is separate from the rest and does not actually take f as an input. For this, we propose directly modelling the feedback with another neural network (the FNet) for use as a loss function. If this FNet is correctly able to learn to copy how we label data with feedback, it could be used as a loss function for regression. Thus, in order to maximize feedback, our loss function would be as follows:

$$Loss_{FNet} = -FNet(s, \hat{\theta})$$

After learning this loss, we can either use it as a loss function to train a policy network, or, every time we want to run inference, we can run a computationally expensive gradient descent optimization to pick the best action. Because the second does not depend on the training distribution (and so we can avoid importance sampling), and just to save on engineering time, we choose an even easier version of the latter: we pick the action predicted by the FNet to be the best, out of a list of discrete options. (These approaches would be analogous to a deep deterministic policy-gradient algorithm and a batch-RL continuous DQN in the RL framework [14] [15]. We do not have to worry about freezing a stale "target" network for the critic DQN because our labels are not changing, and we have no actor because we went with the expensive inference.) One thing to note, is that adding more negative points will not "push" our regression further away from this point, but rather just make our FNet more confident

of the negative feedback there. This may not be the desired effect for all applications. Moreover, the FNet cannot operate on purely positive points with no gradation. That is, behavioral cloning cannot be recovered from it. There may be workarounds that could do this, such as attempting to set a "default" value for the FNet by feeding in random noise as an input and labelling it negatively, but this would likely have little affect, as the noise will never be similar to an input state. (Alternatively, there may be ways to adapt discrete action off-policy maximum entropy Q-networks, such as in NAC [7], to continuous control, such as in SAC [16], to push down the values of actions our networks have not seen; However, for our purposes, we prefer having a singular network, and a network that does not require expensive gradient descent for each inference step, so we did not pursue this after seeing the current unpromising results for our FNet.)

B. Feedback

In order to simply the feedback, we chose to force the feedback to be between -1 and 1. The first data we recorded was 20 minutes of optimal driving and we labelled all of this data with a feedback of 1.0. Choosing the rest of the data, and how to label it, was a bit more tricky.

As a side note, to fit this into the RL framework, we could view this feedback as Q values for either the optimal policy or the current policy (we do not draw a clear distinction since it would not be an easy calculation for our expert labeller to make). Alternatively, we could also view our problem as a contextual bandit, since the feedback for every action falls in the same range. (Although if we take enough bad actions in a row, we will prematurely terminate our episode.)

In order to justify this, we could either view our feedback as analogous to an advantage function of the optimal policy in the RL framework, but to a first approximation, subtracting the mean action-value from the value in the state instead of the maximum action-value, or we could view the problem as a contextual bandit and assume the feedback is the expected reward, if we are just as capable of getting reward in the next state regardless of the current decision. We intended these values in our case to apply to any future policy, as if our problem were a contextual bandit, but we do not want this to be a requirement of our model.

One reason that we are not using reinforcement learning is that letting the car explore actions is dangerous and destructive. In this vein, we wanted to collect data only on "safe" driving. However, the neural network needs data that will tell it "bad" actions and also "good" actions that will recover the car from bad states. In order to accomplish this, we made the goal staying as close to the middle of the road as possible, and labelled any actions that take the car away from the road as negative. Moreover, the speed at which the action takes you away from the center of the road indicates how bad the feedback should be. The opposite is true for positive examples: for positive examples, how fast you return to the center of the road should be how good the feedback is.

In order to explore these types of states and actions, we collected two more types of driving: "swerving" and "lane changing". The first image in Figure 3 below is swerving. In swerving, the car was driving in a "sine wave" pattern on either side of the road. I spent 10 minutes collecting this data on the right side of the road and 10 minutes collecting this data on the left. The second image is lane changing. For this, I drove to the right side of the road, straightened out, staid there, and then returned to the middle. I repeated this for 10 minutes, and then collected 10 minutes on the left-hand side as well.

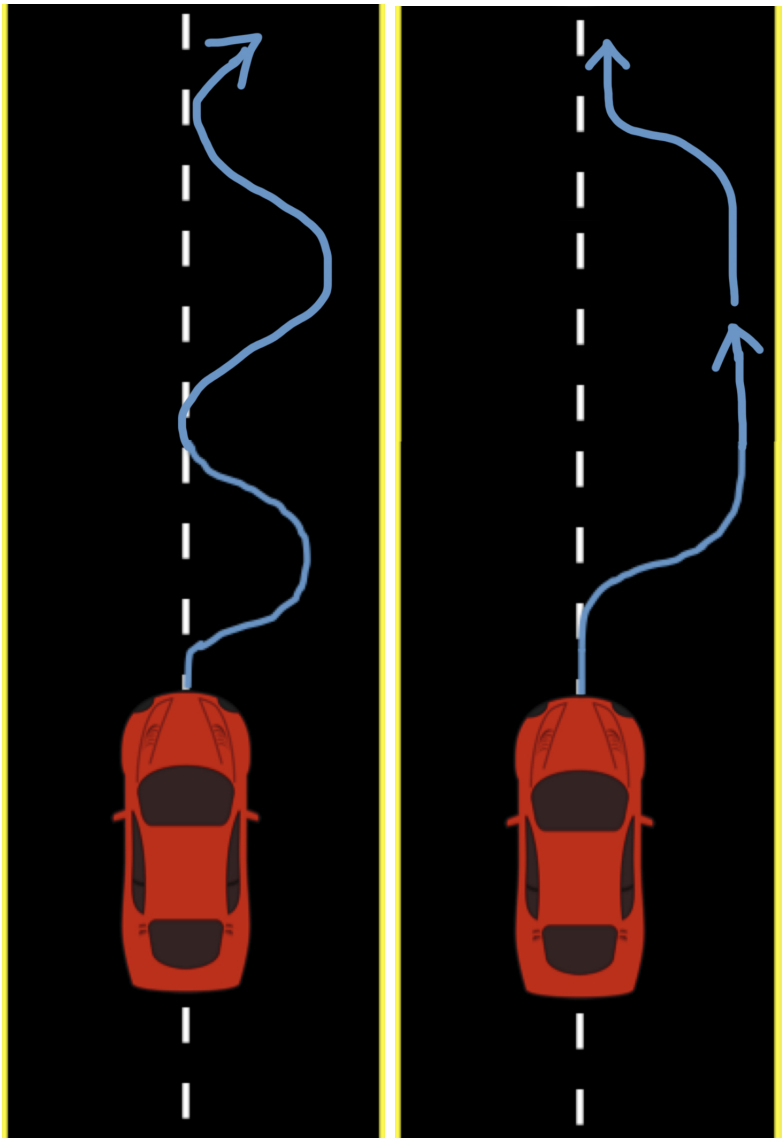

Fig. 3. Types of Driving: Swerving (1st), Lane change (2nd)

At first, we considered labelling all of this data by using a slider that returned values from -1 to 1. However, we realized that this is an issue for two reasons. First, it is very hard to tell when to label the data positively and when to label it negatively, and there are large discontinuity in how you would want the data labelled. For example, it may seem that when swerving away from the road, that this becomes a positive example again when you start heading back towards the center. However, this is not the case: although the state may become "good" again when your are finally heading back towards the middle of the road, the action that actually made that happen (i.e. steering left), began earlier, since steering affects the car over time. Moreover, since heading off the road when you are already near the edge of the road is worse than doing so when you are near the middle of the road (in that it may cause you to crash sooner), when you suddenly start turning left (the point of inflection on the sine wave), you will need to jump to a positive feedback from a very negative feedback, which is a discontinuity that is hard to capture on a slider. And second, there may be small subtleties that are just difficult for humans to capture given their reaction time. For example, when you lane change to the right edge of the road, in order to straighten out, you start to turn left towards the center, which is

positive, but you then turn right again in order to stay straight. This action is hard to capture with a slider.

In order to circumvent all of these issues, we decided to collect feedback using the steering wheel. (Note, for the sake of consistency, we had one person do all of the driving and the other do all of the collecting.) Our first thought was to just turn the steering wheel to the correct angle. However, this is very difficult to do, especially on turns, when you cannot see the actual effects your steering is having on the car. (Also, if we could do this, we could just use this data for more behavioral cloning!) Therefore, what we decided to do instead was to collect differential steering that is proportionally correct. That is, we turned the wheel based on how much more we wanted the car to go in the direction we were steering. This number did not have to be the exact correct angle, it just had to be true that turning the wheel twice as much meant we wanted it to change twice as much.

In order to actually process this data into a +1 to -1 value, we used the following equation below:

FEEDBACK(c, θ)

1 **if** $|c| \geq \theta_{max}$
2 **return** $-\frac{1}{2}$
3 $c = c/\theta_{max}$
4 **if** sign(c) == sign(θ) or $|c| \leq \epsilon$
5 **return** $1 - |c|$
6 **else**
7 **return** $-|c|$

In line 3, we normalize all of our data by diving by a θ_{max} that we pick. (We will return to the case where $|c| \geq \theta_{max}$ later.) So now, by line 4, all of our c values fall between -1 and +1. In line 4, if we are turning the steering wheel in the same direction as the car (with some epsilon of error also being acceptable), then the feedback should be positive. (We set epsilon to $5/\theta_{max}$ so that it allows us to 5 degrees of tolerance.) Remembering that we are recording deltas in steering, a greater delta should result in a less positive signal. Therefore, we subtract c from 1. If we are steering in different directions (line 7), then the feedback should be negative. The greater the delta, the more negative it should be.

One problem with this algorithm occurs on turns. for example, if the car is rounding a left turn, and it is turning left, but not nearly enough, so that it will drive off the road in its current trajectory, we currently have no way to give this a negative value other than by turning the wheel to the right, which is unintuitive. (See Figure 4 below.) Therefore, we added lines 1 and 2 to the code. In these lines, if the magnitude of the correction is large enough, we automatically just map the output to -0.5. It is interesting to note that if the car is off to the right side of the road on a left turn, if the car is still going to make the turn, it will get positive feedback (just not as much as if it turned more left). This is in contrast to what happens on a straightaway, where staying off to the right and going straight results in a negative feedback. None of these seem to be an issue, but it is interesting to note the difference in reward on turns nonetheless. It should also be noted, however, that we never had to make use of the code in lines 1 or 2. Therefore, we simply set θ_{max} to an angle just larger than our largest magnitude. We chose it to be close to our largest magnitude because, if we didn't, all of our feedback would be very close to 0 or 1.

Fig. 4. Intial Feedback Issue with Turns

There are three other interesting thing to note with the feedback. One is that our positive or negative feedback may actually never reach 0 and -1 respectively, since we divided by a constant greater than the largest magnitude. However, our α parameter provides an easy way to scale the relative proportions. Second, slow actions back to the center of the road will be rewarded less than quick actions back to the center of the road. However, this is not true for straightening out at the center of the road. Straightening out at the center will require very little correction and so will have a very positive reward. This was an intentional feature to prevent oscillation of the car about the center of the road. Third, although the car does get more feedback in general for going back quickly, since we are sampling at a constant frame rate and driving at a constant speed of 15 mph, the slow actions will effectively be over-sampled. This does not seem to create any issues in practice, but it is something to be aware of.

C. Architecture

We chose to use only an hour of data so that it was not the case all of our loss functions produced a fully capable and indistinguishable self driving car. (In addition to the sake of time and efficiency.) Also, we sampled states at approximately a rate of 12 frames per second, but then down-sampled further to only keeping every 5th frame, for a result of about 2 frames per second. However, in order to not bias our model towards always turning left and also to get more data, we augmented all of our data by flipping it left-right, inverting the angle label, and leaving the feedback the same. After this augmentation, we had 17,918 training images and 3,162 validation images (a 85:15 split).

One method that helped us to bootstrap learning with limited data was transfer learning. We decided to start with a trained image recognition network. We chose Inception-V3 since, among the top winners of the ImageNet competition, it had few parameters (see Figure 5).

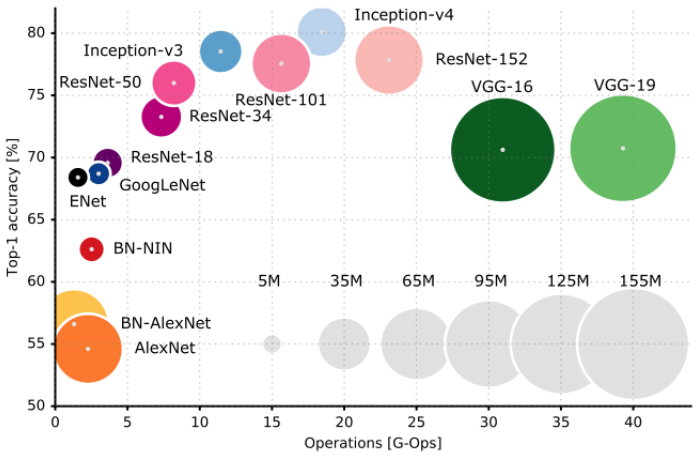

Fig. 5. Inception Parameters vs Others

In order to use Inception Net, we took all of the layers up to some point, and threw out all the layers after that, and then added our own fully connected layers, followed by a tanh activation to ensure that our output was -1 to 1. This -1 to 1 output works both for the PNet and FNet, since our simulator (a modified Udacity simulator coded in Unity) expects a steering value of -1 to 1 and our feedback is also -1 to 1. Note that for our FNet, we also appended 1 additional neuron onto the first dense layer, to be used for the input angle that needs feedback. And our loss for the FNet was mean squared error. We did not add our own convectional layers, since it was easier and better just to allow the gradient to flow through the whole network, changing the weights, and co-opt existing layers that needed to change. It turns out that the "bottleneck" layer that we attached our new fully-connected layers to was one fairly close to the middle. (See the Figure 6.) We assume that this is the case because it can recognize edges and colors and maybe even some basic features of the road at this point, but has not yet moved on to having neurons that represent only high level features such as "cats".

A few hyper-parameters: our fully connected layers had sizes 100, 300, and 20, in order. (101 for our FNet.) Our batch size was 100 and we trained for 5 epochs. Unless otherwise specified for a given model in the results section, we had an α value of 1.0, we did not threshold, and we had a learning rate of 1e-6. As in the Inception model we were using, our input was bilinearly sampled to match the resolution 299x299. Likewise, we subtracted off an assumed mean of 256.0/2.0 and divided by an assumed standard deviation of 256.0/2.0.

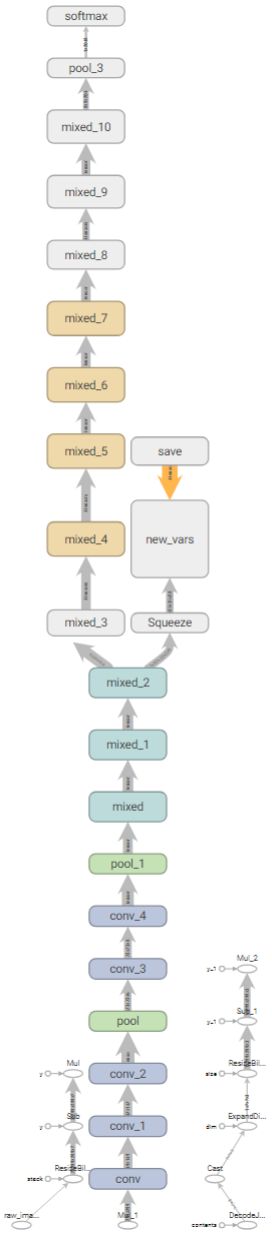

Fig. 6. Our Architecture

IV. RESULTS

A. *Training*

During training, I first tuned our FNet's hyper-parameters. I trained it such that our loss would converge to a low value but not take more than half an hour to train. It happened to be the case that these hyper-parameters worked well for the PNet as well, likely since the losses and architectures are so similar. During training, I kept track of two validation metrics: the loss for the model being trained, and the average absolute error on just the positive data multiplied by 50. The first I refer to as "loss" and the second I refer to as "cloning error" (since it is the 50 times the square root of the cloning error) or just "error". The reason I multiplied by 50 is that this is how Unity converts the -1 to 1 number to a

steering angle, so the error is the average angle our model is off by on the positive data. (This is true with the maximum angle set to 50.) There is a slight caveat on the cloning error: the value is actually a bit lower than the "average absolute error on just the positive data" because we dealt with the negative example loss's by zeroing them out. (But we do this consistently for every model, so it is a great relative metric.) These errors are summarized as a tuple of "(cloning error, loss)", in our plots and reports.

During training, these two metrics generally behaved very similarly, however, in the models for which I increased the learning rate, you can see that these eventually start to diverge. In this case, the error on the positive data started to increase, but the loss was still decreasing. For this reason, I tried varying the learning rate on several models, to see if the loss was more important than the "cloning" error. It is clear that the behavioral cloning models ($\alpha = 0.0$ and THRESH) should in general do better on the "cloning" error, since they are very closely related. Whereas for non-thresholded data, it was trained with examples weighted differently. And for the negative data, it was trained to also get "away" from negative examples. We hope that even though the cloning error may increase, this means that it is because the model is choosing something BETTER than (yet further away from) the positive examples. We still use the cloning error, however, because it is a useful intuitive metric for training and comparison. In figure 7 below, we can see the loss and error diverging for a PNet trained with 10x the normal learning rate.

Fig. 7. Cloning Error vs Loss

Before we discuss how we compared the performance of our models, we should discuss the FNet. Instead of using the FNet as a loss for another neural net, we simply fed in a list of angles and returned the one that our FNet predicted would be best. Since fine granularity in angles is not necessary for this task, by just testing a series of discrete angles, we were better able to see if the FNet would work before committing to an expensive training process. In addition, I would like to mention that, after training, the results of the FNet did not look promising. In fact, what happened is that the FNet learned to predict different feedback for different images, but not for different angles. (See Figure 8 for a random sampling of feedback for given angles rows and images columns.) I suspect that this happened because the

state is a much better predictor of feedback. It is likely enough to predict the feedback based on which way the car is facing, rather than which way the car steers. For example, if the car is heading off road, than pretty much all of those will all have same feedback. The only image of a car driving off the road that will have a different feedback is when the car starts to turn back onto the road. However, this change in angle only lasts a brief amount of time, until the change from the steering actually has an effect and the car has an image state that is not pointing off the road. Only that state where you start to turn back has the same image but a different feedback. For this reason, the FNet's choice was almost never changed, and was almost always the largest or smallest angle possible (depending on stochasticity during training).

```
[[[0.27463108 0.8966497  0.8142399  ... 0.61680347 0.70496565 0.85056734]]
 [[0.27106214 0.8953991  0.8142741  ... 0.6152304  0.70133996 0.8501611 ]]
 [[0.26963258 0.8948949  0.81419414 ... 0.6144671  0.6998794  0.8499993 ]]
 ...
 [[0.25777608 0.8907649  0.8126278  ... 0.6096654  0.68775004 0.84868026]]
 [[0.25635573 0.89021856 0.81237054 ... 0.60916096 0.68617576 0.84817505]]
 [[0.2523774  0.8888414  0.8117536  ... 0.607898   0.68272495 0.8468019 ]]]
```

Fig. 8. FNet Feedback Predictions by Angle (row) and Image (col)

B. Experiments and Benchmarks

We tested our models by running them 8 times in the simulator and recording the time until the car crashed or all 4 tires left the road. In the case that the car drove off the road and came back (which happened very rarely), we discarded that run and started over. It should also be noted that a common failure time was around 0:06 and 1:58. Both of these indicate a failure where the car drove into dirt. There is a patch of dirt near the start that replaces the side of the road, and also a patch of dirt later on that comes right ahead of a sharp turn.

We first tested our PNet with the scalar loss, our PNet with the exponential loss, and our FNet. We plotted their mean times over 8 runs and their standard deviations. See Figure 9. (All of these were on the default settings, except for the exponential loss model, since we could not get many exponential models train correctly.)*

*All of these were on the default settings, except for the exponential loss model. The default exponential model did not converge during training. Instead, we chose to set α to 0.1, which allowed the model to converge and had a loss summarized by the tuple (error=4.978, loss=0.1544). Whereas, the default exponential model was off by an average of 29.03 degrees, with a loss tuple of (29.03, 32.25). Other variation tested included training for 11 epochs (11.24, 0.03817), THRESHOLD (32.316, 32.32), α=0.1 (diverged and not recorded), α=0.1 THRESHOLD (5.876, 0.1669).

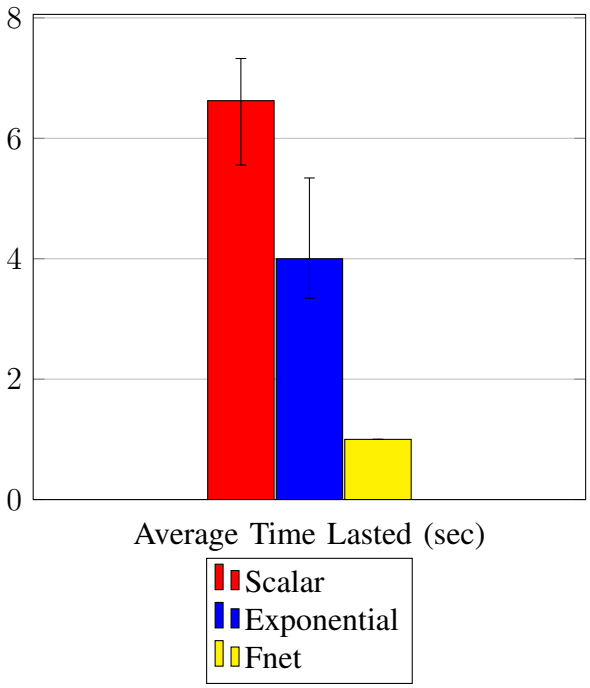

Fig. 9. Initial Comparison of Loss Functions.

Given that the Scalar loss performed best (and was training correctly), we spent more time creating and bench-marking variations of this loss function but with different hyper-parameters. This can be seen in Figure 10. Figure 11 underneath will remind you of the default parameters if not specified.

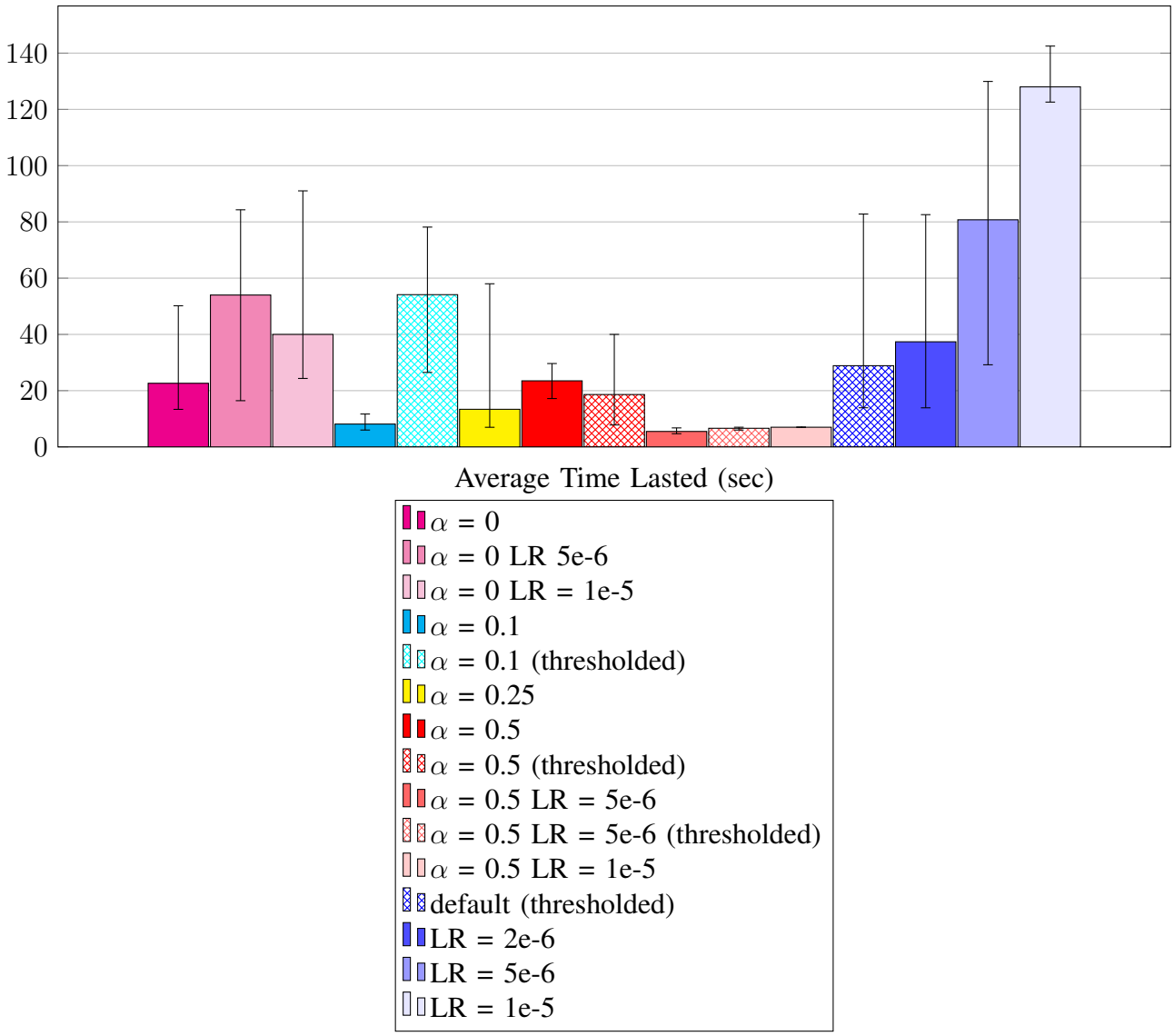

Fig. 10. Scalar Loss Function Performance by Hyper-Parameter

Name	Default Value	Description
LR	1e-6	Learning rate
Thresholded	False	If True, all feedbacks are thresholded to 1 or -1
α	1.0	How much to scale negative feedbacks down by, in [0, 1]

Fig. 11. The default hyperparameters we used for the scalar PNet.

From this plot we can learn several interesting things. The first interesting thing to note is that with thresholding the feedback to -1 or +1 (the blue crosshatch pattern), the scalar performance came up to about 30 seconds (compared to the 6 seconds from our previous plot). Although at first this could be taken to indicate that having gradations in positive and negative data could be harmful, we can see that the same performance can be achieved by increasing the learning rate instead of thresholding, by looking to the three blue bars to the right. I believe the reason for this is simply that we tuned the learning rate to work well on thresholded data, and so, when we don't threshold or clone our data, the scalar on the loss drops significantly, forcing the network to take smaller steps, and effectively

decreasing the learning rate. (For this reason, the next plots we look at will be cross-validation plots in order to tune the learning rate.)

The second interesting thing to note is the pink group. In the pink group, all of the α values are 0. In particular, the last two pink bars are identical to the last two blue bars, except for the fact that the alpha value is 0. This indicates that, not only are gradations in the data useful, but also that throwing away the negative data is harmful.

Overall, the best model from these experiments was the default Scalar settings but with a learning rate that was 10 times greater. Moreover, on individual trials, it is worth noting that the Scalar loss with 5 and 10 times the learning rate had the two longest runs out of any model (155 and 143 seconds respectively) and the Scalar loss with 10 times the learning rate is only model that scored over 120 seconds multiple times, let alone every time.

In order to evaluate how the scalar loss compares to behavioral cloning, we revisited the learning rates of each in figure 12 and 13, respectively.

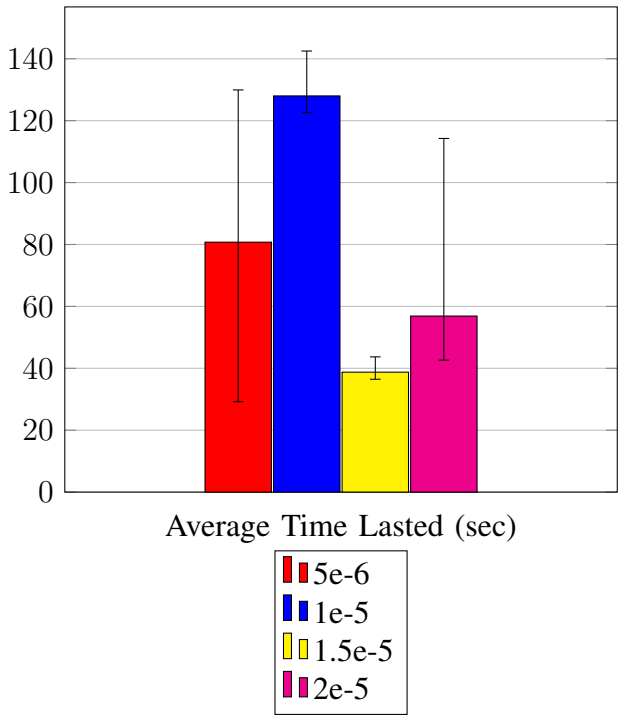

Fig. 12. The scalar loss performed best with a learning rate of 1e-5.

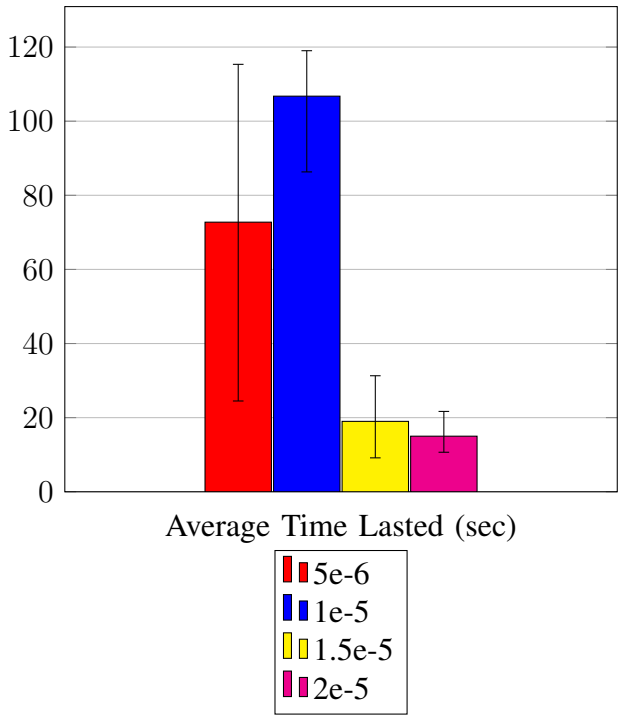

Fig. 13. The behavioral cloning loss performed best with a learning rate of 1e-5.

After selecting the best learning rate for each model (which turned out to be 10 times the default learning rate, for both of them) we then trained new versions of this network 2 more times. Each time, we ran it for 8 runs and calculated the performance as the mean over these 8 runs. We then calculated the mean performance over these 3 training sessions. We compared this to the mean over 3 training sessions of the average performance of a behavioral cloning network. Figure 14 shows the results.

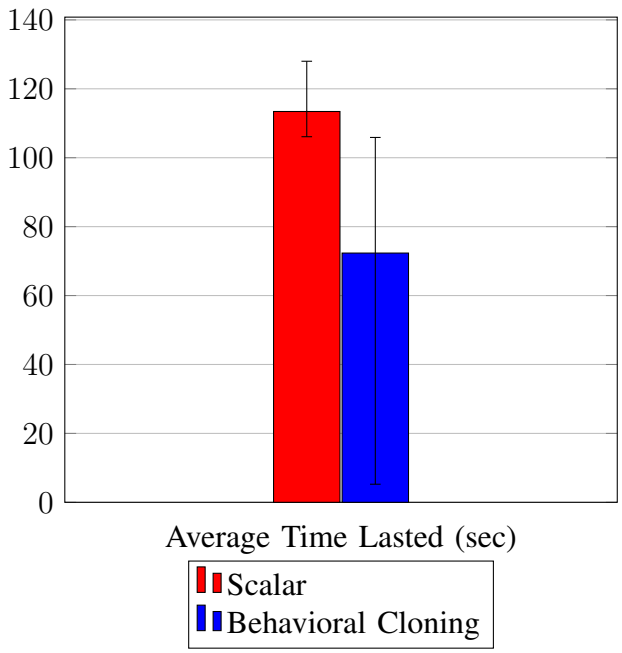

Fig. 14. Our policy net using feedback as a scalar in the loss function worked better than behavioral cloning.

As you can see, over 8 runs and over 3 separate training sessions, our best model trained by a Scalar loss performed over 1.5 times as well as the standard behavioral cloning benchmark, with significantly less variance. (In fact, the lower error bar for our scalar loss is approximately at the same value as the upper error bar for our behavioral cloning loss.)

V. CONCLUSION

We hypothesized that, for our autonomous vehicle setting, adding in negative examples would allow our models to have better performance. We believed that one of our loss functions would accomplish this task. Given that our scalar model performed over twice as well as the behavioral cloning model, I believe that we have shown this to be true. The specific method of regression with negative examples we used allows for learning deterministic continuous control problems from demonstration, with the ability to recover behavioral cloning on only positive examples. Moreover, the loss function that empirically worked the best in this domain does not require an additional neural network to model it, and it induces a stochastic policy gradient that could be used for fine-tuning with RL. We thus believe ReNeg could be useful in the autonomous control industry: with no additional real world time, increased performance in supervised learning can be achieved by simply labelling the data as it is collected, and then RL or additional supervised learning could be used thereafter.

VI. KNOWN ISSUES

Due to our simulator being down, we tested on a different computer than we trained on. Because of this, our graphics settings were left slightly different at test time as compared to training. That is, we ran in Unity editor instead of the build (with the specific graphics settings), and we left the resolution being sent to the PNet as 320x160 (instead of 640x320). However, the images all are re-sized to 299x299 when they are processed for the neural net, so we really only lost vertical resolution. We expect our results to be the same relative to each other and to be very similar to if we had run at the appropriate settings. Running it now, It seems that our models perform a little bit better, but very close

to how they did when we benchmarked. Certainly our best model is still better than our benchmark. However, it may be useful to redo these benchmarks on the correct settings to show whether the neural networks that performed better did so due to greater tolerance of small resolution changes.

Additionally, our car tends to swerve on the road to various degrees. This may be due to a mislabelling of our serving data. When we initially labelled our data, we set an epsilon of 5 degrees, meaning that any correction of less than 5 degrees was mapped to a correction of 0, resulting in a perfect feedback of 1. It is possible we collected with an epsilon that was too large and the beginning of swerves was encouraged.

Finally, we realized after the fact that we actually labelled our data at a speed a bit slower than real time. If we actually want to test that we can do the labelling simultaneously with the collection of the driving data, then we would have to relabel the data. However, since this speed only affects how much skill is required in the labelling, and since we did not have much practice before labeling the data, we are confident that it could be done slightly faster at real time.

VII. FUTURE RESEARCH

Future research should focus on 2 things: the loss function and enforcing continuity.

A. Loss Function

Immediate next steps should likely focus on alterations to the loss function; there do exist ways we can achieve all desired properties 1) to 4). There are issues with both our scalar loss function and our exponential loss function. Our exponential loss function was created to satisfy property 4): the magnitude of the loss for negative examples should drop off exponentially with respect to the distance. This is a problem since, in our scalar loss, the negative examples actually have an exponentially increasing affect as the distance increases. Although the exponential loss accomplishes this solution, there is a dilemma: our exponential solution violates the desired the desired property 3) as mentioned earlier. That is, we want the magnitude of the loss to increase with the magnitude of the feedback, f. However, for positive examples, this only is the case when the absolute difference between θ and $\hat{\theta}$ is greater than 1, and for negative examples, this is only the case when the absolute difference is less than 1. Although this may not be a large concern for the negative examples, since the difference is at most 2, this could be an issue for our positive examples. This issue with the exponential loss function can actually be solved in two different ways. First, I can think of an elegant solution that will modify our scalar loss to have this exponential decay property and satisfy all desiderata. Second, I an think of a way to modify our scalar loss to have neither an exponential increase with distance nor an exponential increase. And third, I can think of an "ugly" patch for our exponential function, if we really must have the loss be exponential in f.

1. We can accomplish this exponential decay by modifying our scalar function in a very easy way: Move the sign of f into the exponent:

$$Loss_{scalar} = |f| * (\theta(s) - \hat{\theta}(s))^{2*sign(f)}$$

Using this loss function we have all three properties satisfied. That is, positive examples encourage moving towards them, negative examples encourage moving away from them, and the amount of this movement increases with the magnitude of f. Moreover, we also have the properly that, in negative examples, loss drops of exponentially with the distance from the negative example (because we are dividing by it).

2. If we want our scalar loss function to have neither an exponential decay nor an exponential increase with the distance from the negative points, we can simply use the following loss:

$$Loss_{scalar} = f * |\theta(s) - \hat{\theta}(s)|$$

This has the not-so-nice property that, in the positive example, it allows outliers much more easily than the traditional squared loss. However, it has the very nice property that, given a single state input, as long as you have more positive examples than negative examples, your loss will always be minimized in that state by a value between your positive examples. This is because, as soon as you get to your greatest or least positive example, every step away from your positive examples will cost you 1 loss, for each positive example you have, and you will only lose 1 loss for each negative example you have. (Note, if you are not thresholding, then this translates to more total $|f|$ for positive examples than negative examples.)

3. If we really want our feedback to be in the exponent, we can accomplish this with a more complex solution. We can split our examples into positive and negative examples and have a loss for each. The positive and negative loss functions could respectively be:

$$Loss_{exp,positive} = min(|\theta(s) - \hat{\theta}(s)|/tol, 1)^{2f}$$

$$Loss_{exp,negative} = (|\theta(s) - \hat{\theta}(s)|/2)^{2f}$$

For the positive examples, the magnitude of the loss increases with f in the case that the number being raised to the $2f$ is greater than 1. If $|\theta(s) - \hat{\theta}(s)|$ is greater than tol, we have both ensured this is the case and we are back to our original loss function, but scaled by a constant factor of tol. But what happens if the difference is less than tol? In that case, for any estimate $\hat{\theta}$ within tol of the intended $theta$, our parameters will have a derivative of 0 (no effect), due to the min. (If we want this to be in degrees, we can just set adjust this to tol/=50). (This will cause the neural network to forget about examples that it does well on until they become sufficiently problematic again. This may be helpful in that the network can focus on the area it needs to, or it could be harmful in that is prevents convergence.) For the negative examples, the magnitude of the loss will only increase with f if the number being raised to the $2f$ is less than 1, because we are dividing my it. To ensure this is the case, we can just divide by our range, which is 2.0. To combine these two functions, we can just pick the best convex combination of them.

B. Continuity

Because, in both our scalar and exponential loss, our loss function at a given state with just a negative example is minimized by moving away from the negative example, our regression in that state will tend toward positive or negative infinity. Certainly having a cost on negative examples that drops of exponentially will help, but it may not be enough. Moreover, we may not want to rely on the structure of neural networks to discourage this discontinuity. Therefore, research could be done on adding a regularization term to the loss that penalizes discontinuity. That is, we would add some small loss based on how dissimilar the answers for nearby states are. Of course, this implies a distance metric over states, but using consecutive frames may suffice.

REFERENCES

[1] S. Ross, G. J. Gordon, and J. A. Bagnell, "No-regret reductions for imitation learning and structured prediction," *CoRR*, vol. abs/1011.0686, 2010.

[2] S. Ross and J. A. Bagnell, "Reinforcement and imitation learning via interactive no-regret learning," *CoRR*, vol. abs/1406.5979, 2014.

[3] W. Sun, A. Venkatraman, G. J. Gordon, B. Boots, and J. A. Bagnell, "Deeply aggrevated: Differentiable imitation learning for sequential prediction," *CoRR*, vol. abs/1703.01030, 2017.

[4] T. Degris, M. White, and R. S. Sutton, "Off-policy actor-critic," *CoRR*, vol. abs/1205.4839, 2012.

[5] C. J. C. H. Watkins and P. Dayan, "Q-learning," in *Machine Learning*, pp. 279–292, 1992.

[6] R. Munos, T. Stepleton, A. Harutyunyan, and M. G. Bellemare, "Safe and efficient off-policy reinforcement learning," *CoRR*, vol. abs/1606.02647, 2016.

[7] Y. Gao, H. Xu, J. Lin, F. Yu, S. Levine, and T. Darrell, "Reinforcement learning from imperfect demonstrations," *CoRR*, vol. abs/1802.05313, 2018.

[8] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *In Proceedings of the Twenty-first International Conference on Machine Learning*, ACM Press, 2004.

[9] K. Shiarlis, J. Messias, and S. Whiteson, "Inverse reinforcement learning from failure," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, (Richland, SC), pp. 1060–1068, International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[10] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. I–387–I–395, JMLR.org, 2014.

[11] N. Meuleau, L. Peshkin, L. P. Kaelbling, and K. eung Kim, "Off-policy policy search," tech. rep., 2000.

[12] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *CoRR*, vol. abs/1604.07316, 2016.

[13] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, pp. 229–256, May 1992.

[14] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *CoRR*, vol. abs/1509.02971, 2015.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, "Playing atari with deep reinforcement learning," *CoRR*, vol. abs/1312.5602, 2013.

[16] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *CoRR*, vol. abs/1801.01290, 2018.