<table>
<tr><td>Title</td><td>Model Free Imitation Learning with Policy Optimization
Jonathan Ho HOJ@CS.STANFORD.EDU</td></tr>
</table>

<table>
<tr><td>Abstract</td><td colspan="3">> Existing imitation learning algorithms involve solving a sequence of planning or reinforcement learning problem, and it is limited to be directly applicable to large high dimensional environments, and the performance can significantly degrade if the planning problems are not solved to optimality.
> The method is based on policy gradients, and under the formalism of apprenticeship learning. Meanwhile, it is also a model-free algorithms.
> The approach scales to large continuous environments.</td></tr>
<tr><td rowspan="4">Introduction</td><td>Methods</td><td>Advantages</td><td>Disadvantages</td></tr>
<tr><td>The simplest approach to imitation learning is behavioral cloning.</td><td>Conceptually simple and theoretically sound</td><td>Cascading errors</td></tr>
<tr><td>Inverse Reinforcement Learning (IRL)</td><td>One of the most successful approaches to imitation learning, assume the behavior the learner desire to imitate is generated by an expert behaving optimally with respect to an unknown cost function. Do not suffer from cascading error actions.
IRL generalize expert behaviors.</td><td>Great expensive</td></tr>
<tr><td>Contribution</td><td colspan="2">Develop a gradients based optimization formulation over parameterized policies for apprenticeship learning.
Propose two model free realization of these optimization algorithms: standard policy gradients algorithm, and policy gradient algorithm that incorporates trust region constraints to stabilize optimization.</td></tr>
<tr><td>Preliminaries</td><td colspan="3">>basic notation from reinforcement learning

> Stationary stochastic policies
State action value
state visitation distribution</td></tr>
<tr><td>Apprenticeship learni</td><td>Basic idea</td><td colspan="2">> Apprenticeship Learning is that the learner must find a policy that performs at least as well as the expert.
> The Apprenticeship Learning algorithm try to find a policy that minimizes the</td></tr>
</table>

ng		cost difference of policy computed from AL and expert policy. Having defined the objective, the job of an apprenticeship learning algorithm is to solve the optimization problem $$\underset{\pi}{\text{minimize}}\ \delta_{\mathcal{C}}(\pi, \pi_E). \qquad (3)$$ > Two ingredients must be provided in order to instantiate the frame: cost function class (assume the cost function belongs to certain cost function), and optimization algorithm to solve the problem.
	AL example	**Feature expectation matching:** Define the cost class as a certain set of linear combination of these basic function: $$\mathcal{C}_{\text{linear}} = \left\{ c_w \triangleq \sum_{i=1}^{k} w_i c_i \,\middle\|\, \|w\|_2 \le 1 \right\}$$ solve the function by inverse reinforcement learning. **Game-theoretic approach:** proposed two AL algorithm: MWAL and LPAL $$\mathcal{C}_{\text{convex}} = \left\{ c_w \triangleq \sum_{i=1}^{k} w_i c_i \,\middle\|\, w_i \ge 0,\ \sum_i w_i = 1 \right\}$$ The weight constrains on the basis functions to line on the simplex allows the maximization over costs to be performed instead over a finite set.
Policy optimization for AL		The gradient-based optimization include two procedures, 1) fitting a local reinforcement learning problem to generate learning signal for imitation. 2) improving the policy with respect to this local problem.
	Policy gradient	>Use the stochastic gradient descent >Duo to the high variance incurred by stochastic gradient descent, the author adopt *Trust Region Policy Optimization*, a model free policy search algorithm capable of quickly training large neural network stochastic policies for complex tasks.
	TRPO for RL	The Vanilla policy gradient methods to improve the policy: $$\eta(\pi) = \eta(\pi_0) + \mathbb{E}_{\rho_\pi} \mathbb{E}_{a \sim \pi(\cdot\|s)} [A_{\pi_0}(s,a)]$$ $$L(\pi) \triangleq \eta(\pi_0) + \mathbb{A}_{\pi_0}(\pi)$$ The problem lie in the step size, how to improve the step size? $$M(\pi) \triangleq L(\pi) + \frac{2\epsilon\gamma}{(1-\gamma)^2} \max_s D_{\text{KL}}(\pi_0(\cdot\|s) \,\|\, \pi(\cdot\|s))$$ The optimization problem finally converted into $$\underset{\theta}{\text{minimize}}\ L(\pi_\theta) \quad \text{s.t.} \quad \overline{D}_{\text{KL}}(\pi_0 \,\|\, \pi_\theta) \le \Delta$$

	TRPO for AL	$\begin{array}{ll} \underset{\theta}{\text{minimize}} & \sup_{c \in \mathcal{C}} L^c(\pi_\theta) - \eta^c(\pi_E) \\ \text{subject to} & \overline{D}_{\mathrm{KL}}(\pi_0 \parallel \pi_\theta) \leq \Delta \end{array}$	
Experiments	Evaluated the approach in a variety of scenarios: finite gridworls of varying sizes, the continuous planar navigation task, highway driving simulation		
Future Work	Generative adversarial networks (Goodfellow et al., 2014), the policy parameterizes a generative model of state-action pairs, and the cost function serves as an adversary.		