1502.05477 - Trust Region Policy Optimization

- **Author: Yunqiu Xu**
- Other reference:
  - https://zhuanlan.zhihu.com/p/26308073
  - https://zhuanlan.zhihu.com/p/29918825
  - https://zhuanlan.zhihu.com/p/30548114
  - http://yixinlin.net/trpo/presentation.pdf
- 这里TRPO就当了解下, 主要看之后的PPO

- Policy optimization
  - Policy iteration
  - Policy gradient
  - Derivative-free optimization: e.g. cross-entropy
- From PG to TRPO:
  - PG

```
Initialize policy π|θ
while gradient estimate has not converged do
    Sample trajectories using π
    for each timestep do
        Compute return and advantage estimate
    end for
    Refit optimal baseline
    Update the policy using gradient estimate ĝ
end while
```

  - TRPO

```
while gradient not converged do
    Collect trajectories (either single-path or vine)
    Estimate advantage function
    Compute policy gradient estimator
    Solve quadratic approximation to L(π_θ) using CG
    Rescale using line search
    Apply update
end while
```

- One weakness of PG is that it's crucial to choose step size
- In TRPO, minimizing a certain surrogate objective function guarantees policy improvement with non-trivial step sizes

- TRPO:
  - 使用single path或者vine方法收集一系列的状态-动作序列，然后使用Monte Carlo方法得到Q值
  - 根据得到的Q值估计出 L_θ 的近似值
  - 在满足散度在一定范围内的条件下更新参数 θ，更新参数时使用了共轭梯度和线性搜索
- 和后面PPO对比的不足
  - TRPO需要满足约束保证散度在一定范围内, 这个过程比较复杂而且不易泛化
  - PPO直接把约束放在loss function里面, 既要防止走了不正确的方向, 又要防止步伐太大扯着蛋
  - PPO结构和速度比TRPO简单, 且泛化性能更好