

**Introduction**

- Likelihood ratio policy gradient methods are state of the art techniques for reinforcement learning in continuous state spaces.

- Learning to hit balls with a bat [7].

- Model-free learning with strong convergence guarantees.

- Show how policy gradient methods can be derived from an importance sampling perspective.

- Show more general form of optimal baselines.

- Present a new policy search method which leverages these insights to outperform standard likelihood ratio PG methods.

**Main Result: IS and Policy Gradients**

\[
U(\theta) = \sum_{t=1}^{T} \log \pi(a_t|s_t) \sum_{s_t} \sum_{a_t} R(s_t, a_t) \pi(a_t|s_t)
\]

- The sample estimate of the gradient of \(U(\theta)\) evaluated using only sample trajectories drawn under \(\pi\) is equal to the likelihood ratio based sample estimate of the gradient of \(U(\theta)\).

- Suggests LRPG does not make full use of data.

**Algorithm Summary**

- **Input:** domain of policy parameters \(\theta\), initial policy \(\pi_0(\theta)\).
  - for \(i = 0 \ldots d - 1\) do
    - Run \(i\) trials under policy \(\pi^{(i)}\), collect returns \(R_i\).
  - while (\(R_i\) is improving) do
    - Find \(\pi^{(i)}\) that maximizes \(R_i\).
  - end while

- **Optimal policy:** \(\pi^* = \arg \max_{\pi} \mathbb{E}_\pi [R]\)

**Main Result: Generalized Baselines**

- For any distribution \(P(X)\), any scalar valued function \(g(X)\), and any bias vector \(b\):

\[
\mathbb{E}_P[g(X)] = \mathbb{E}_P[b^T \nabla \log P(X) g(X)]
\]

- Our approach: find local optima of \(U(\theta)\) through memory-based optimization.

- Use minimum variance base-line for estimating \(U(\theta)\).

- Minimum variance is also an expectation: in principle, can rebase base-line trivially.

- Introducing baselines increases model complexity. Requires more samples.

- Use effective sample size (ESS) to limit search areas of \(t\)-space with many samples.

- Do optimal time search (Armijos rule) [2].

**Proposition:**

- For any distribution \(P(X)\), any scalar valued function \(g(X)\), and any bias vector \(b\):

\[
\mathbb{E}_P[g(X)] = \mathbb{E}_P[b^T \nabla \log P(X) g(X)]
\]

**Proposition:**

- For any distribution \(P(X)\), any scalar valued function \(g(X)\), and any bias vector \(b\):

\[
\mathbb{E}_P[g(X)] = \mathbb{E}_P[b^T \nabla \log P(X) g(X)]
\]

- Use minimum variance base-line for estimating \(U(\theta)\).

- Minimum variance is also an expectation: in principle, can rebase base-line trivially.

- Introducing baselines increases model complexity. Requires more samples.

- Use effective sample size (ESS) to limit search areas of \(t\)-space with many samples.

- Do optimal time search (Armijos rule) [2].

**References**


