Deep Recurrent Optimal Stopping

Niranjan Damera Venkata,

Digital and Transformation Organization,

HP Inc., Chennai, India

Chiranjib Bhattacharyya,

Dept. of CSA and RBCCPS,

Indian Institute of Science, Bangalore, India

INDIAN INSTITUTE OF SCIENCE

भारतीय विज्ञान संस्थान

NeurIPS 2023 , New Orleans

Optimal stopping not well-developed in non-Markovian settings

What is the optimal time to exercise a stock option?

- This is an **optimal stopping problem**
- Typically solved in the **restrictive Markovian setting** invoking the efficient market hypothesis
- State of the art methods are based on deep neural networks (DNNs)

This work explores **model-free** optimal stopping algorithms effective for **non-Markovian** settings, leveraging recurrent neural networks (**RNN**s).

Non-Markovian settings pose fundamental challenges!

Curse of dimensionality:
Explosion of augmented state and parameter space

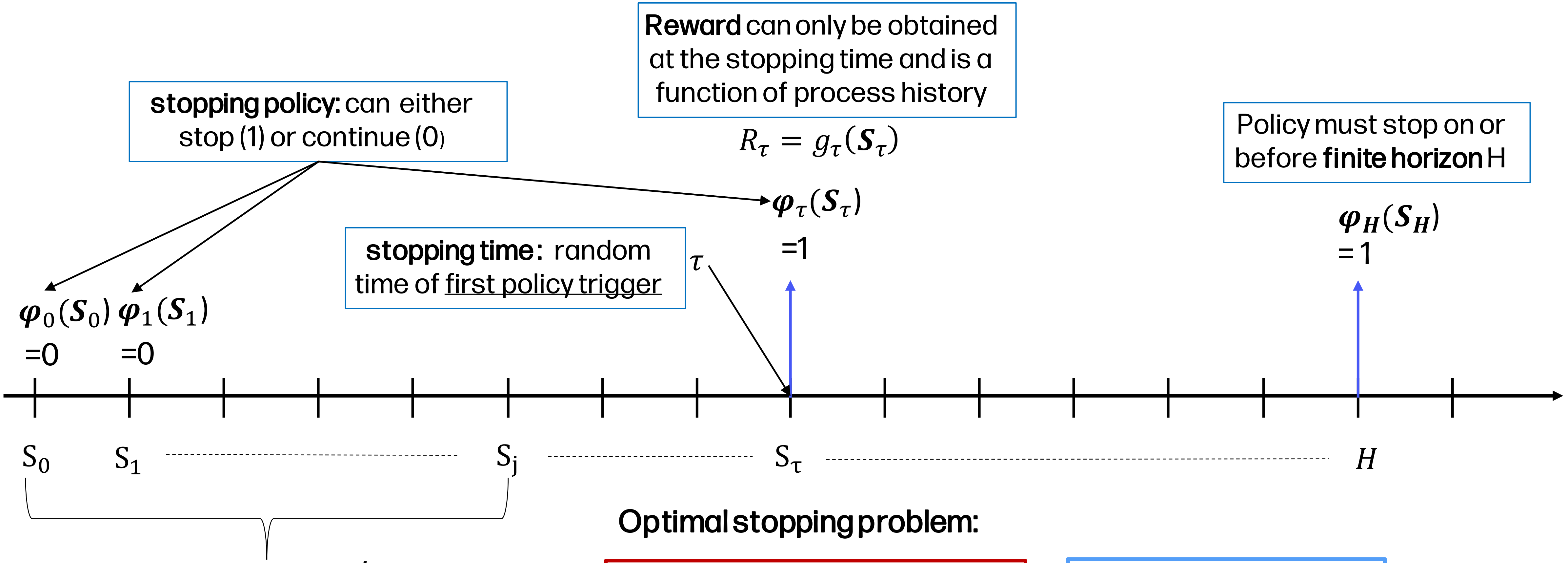
Suitable parameterization of state space
(e.g., using RNNs)

Curse of non-Markovianity:
recursive value estimation algorithms are not suitable

Explore direct policy learning methods
(e.g., policy gradients)

Non-Markovian optimal stopping problem

we consider the discrete-time finite-horizon case



$$\boldsymbol{S}_j = \{S_k\}_{k=0}^{j}$$

process history up-to time step j

Optimal stopping problem:

Solve for τ^* such that:

$$\mathbb{E}[R_{\tau^*}] = \sup_{0 \le \tau \le H} \mathbb{E}[R_\tau]$$

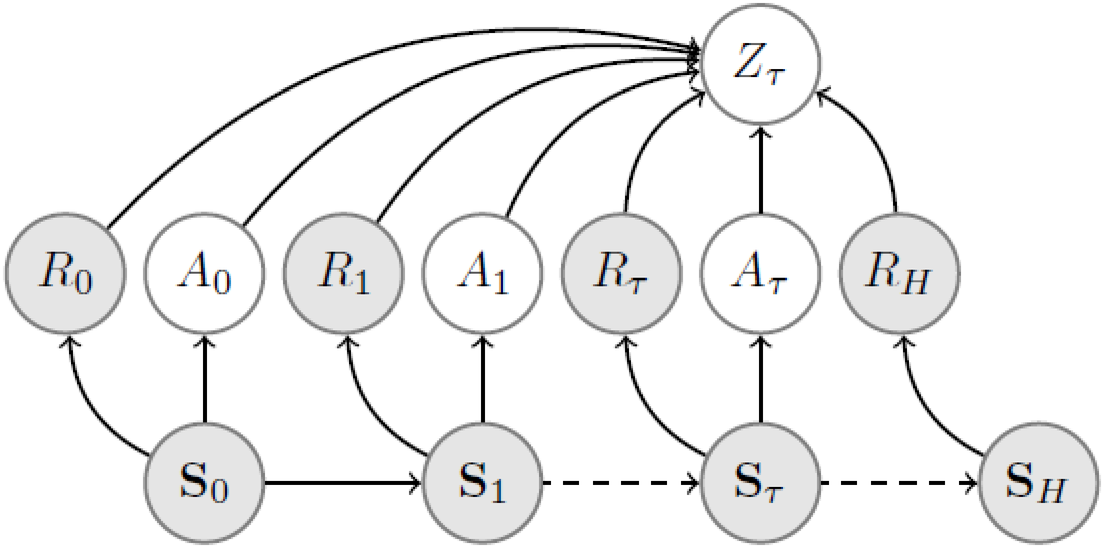
Markovian case:

$\{S_k\}$ is a Markov process

$R_\tau = f_\tau(S_\tau)$

Bayes net reward augmented trajectory model (**RATM**)

represents non-Markovian state-action-reward trajectories



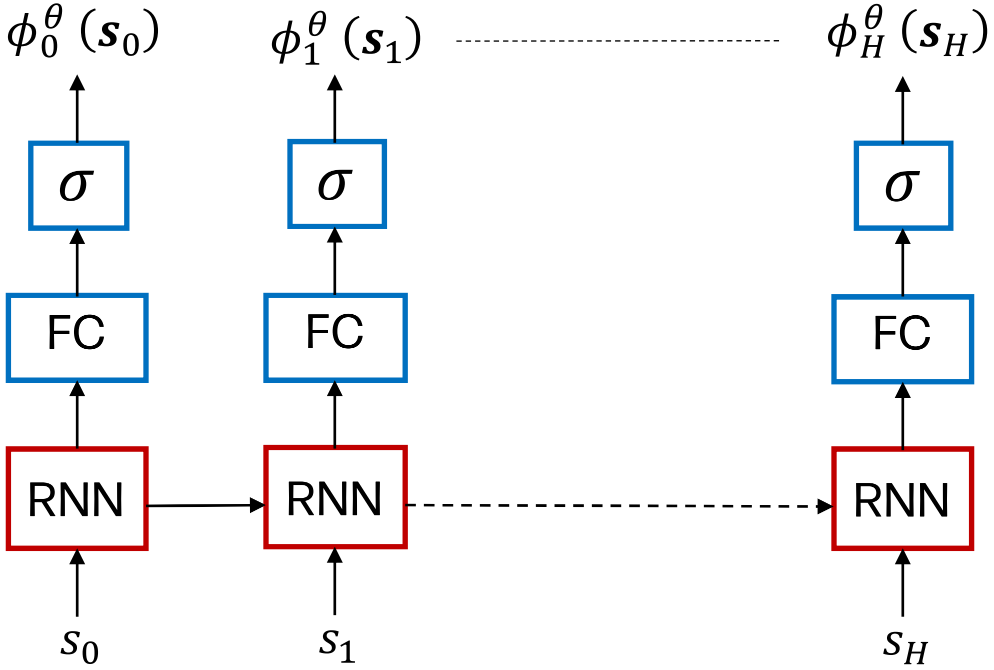
at time step j:

$\boldsymbol{S}_j$: **process history**

A_j: {0,1} **policy actions**

R_j: **reward achievable**

Z_j: {1,0}, 1 if reward is obtained when $\tau = j$



$$\mathbb{P}(A_j = 1 \mid \boldsymbol{S}_j) := \phi_j^\theta(\boldsymbol{S}_j)$$

stochastic stopping policy $\phi_j^\theta(\boldsymbol{S}_j)$ can be **parameterized by an RNN** preventing state and parameter space explosion.

Inference over **RATM** leads to direct policy optimization

$Z := Z_0 \oplus Z_1 \oplus \cdots \oplus Z_H$ (↑ XOR)

Binary RV $Z = 1$ if reward is obtained over a trajectory

$\mathbb{P}(Z = 1 | \boldsymbol{R}_H, \boldsymbol{S}_H; \theta)$ obtained via **inference on RATM**

↓ **RATM inference**

$$J_{WML}(\theta) = \sum_{i=1}^{N} \widetilde{w}_i \log \mathbb{P}(Z = 1 | \boldsymbol{r}_{iH}, \boldsymbol{s}_{iH}; \theta)$$

↑ **trajectory weights**

↓ **E-M**

θ^* **optimal policy**

Bayes net inference leads to **direct policy optimization**, <u>mitigating the curse of non-Markovianity</u>

Optimal stopping policy gradients (OSPG)

offline policy gradient algorithm that eliminates Monte Carlo policy rollouts

Claim (OSPG): ***Incremental E-M*** *with a single gradient step instead of full M-step* ***is equivalent to a policy gradient method***

OSPG highlights

- **First policy gradient algorithm** for optimal stopping
- **Offline** algorithm without **expensive Monte Carlo policy rollouts**
- **Advantage over E-M** is that it can be implemented with **SGD.**
- Optimizes value functions **without recursion**

Optimal Stopping Policy Gradient (OSPG)

$$\nabla_\theta J_{OS}(\theta) = \mathbb{E}_{\boldsymbol{s}_H \sim \mathbb{P}(\boldsymbol{s}_H)}\left[\sum_{j=0}^{H} r_j \psi_j^\theta(\boldsymbol{s}_j) \nabla_\theta \log \psi_j^\theta(\boldsymbol{s}_j)\right]$$

works with **offline process trajectories**

Bayes net inference is used to **eliminate expensive Monte Carlo policy rollouts**

Relationship of OSPG with Value function based methods

Claim (OSPG and Value functions): *OSPG can equivalently be expressed using empirical stopping and continuation values*

Value form of OSPG

$$\nabla_\theta J_{OS}(\theta) = \mathbb{E}_{\boldsymbol{s}_H \sim \mathbb{P}(\boldsymbol{s}_H)}\left[\sum_{j=0}^{H}\left\{\frac{v_j\left(1-\phi_j^\theta(\boldsymbol{s}_j)\right) - k_j \phi_j^\theta(\boldsymbol{s}_j)}{\phi_j^\theta(\boldsymbol{s}_j)\left(1-\phi_j^\theta(\boldsymbol{s}_j)\right)}\right\}\nabla_\theta \phi_j^\theta(\boldsymbol{s}_j)\right]$$

v_j: empirical **stopping value**

k_j: empirical **continuation value**

calls for increasing stopping probability if:

empirical **ratio of stopping value to continuation value** $\dfrac{v_j}{k_j} > \dfrac{\phi_j^\theta(\boldsymbol{s}_j)}{1-\phi_j^\theta(\boldsymbol{s}_j)}$ **odds of stopping** under the current policy

Empirical evaluations on computational finance benchmarks

Experiments in financial derivative pricing

- Pricing Bermudan max-call options
- Pricing American geometric-average call options
- Pricing non-Markovian financial derivatives

OSPG performs competitively with state-of-the-art option pricing methods even in Markovian settings while outperforming in non-Markovian settings!

More results and details in the paper.

Thanks!