SwiftTD: Fast and Robust Temporal Difference Learning

Khurram Javed, Arsalan Sharifnassab, Richard S. Sutton

The Prediction Problem

- Agent receives Φ_t and γ_t at time t
- It predicts a scalar v_t
- It lives for T time steps
- Its performance is evaluated as:

$$\mathcal{L}(T) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T}\left(v_t - \sum_{j=t+1}^{\infty}\gamma_j^{j-t-1}\boldsymbol{\phi}_j[i]\right)^2,$$

where **Φ[i]** is a scalar called the *cumulant* (*e.g.,* the reward)

Limitations of Current Methods

- Learning with large step-size parameters is unstable and diverges
- Learning with small step-size parameters is sample inefficient
- Replaying same data multiple times is computationally wasteful

Proposed Solution: SwiftTD

SwiftTD enables effective learning from large step-size parameters. It has three components:

1) Step-size optimization [1]

Meta-learn the per-feature step-size parameters using a computationally efficient gradient-based meta-learning algorithm

2) Bound on the rate of learning

Bound the rate of learning to make the learner robust by enforcing:

$$1 < \sum_{i=1}^{n}\alpha\boldsymbol{\phi}_t[i]^2 < 0.$$

Requires using True Online TD(**λ**)

3) Step-size decay

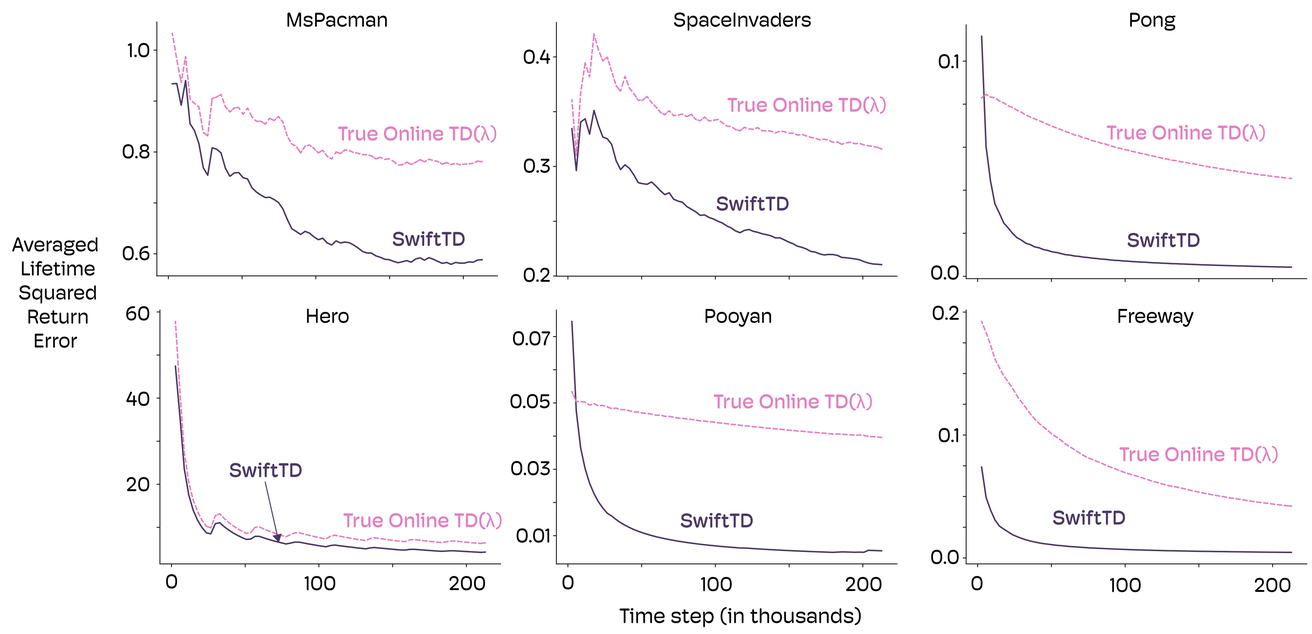
Decay the step-size parameters when the rate of learning is too large

The Atari Prediction Benchmark [2]

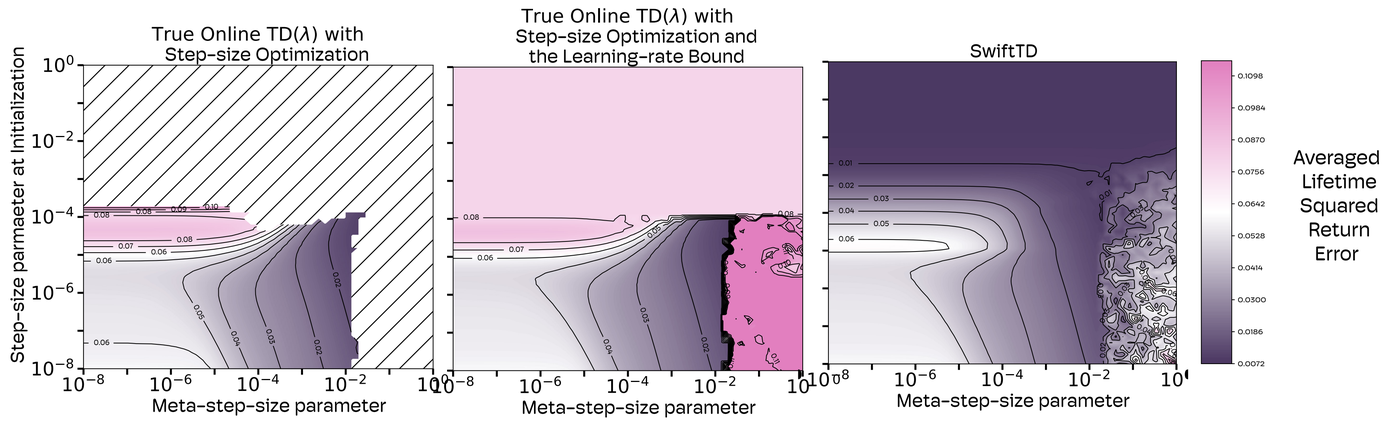
- Learn value function for pre-trained raindbow-DQN policies
- Use 30 minutes of gameplay data
- Pre-process the game frame to get ~200k features (a lossy one-hot coding)
- Clip rewards to be in [-1, +1]
- Use lifetime prediction error for evaluation

Empirical Evaluation

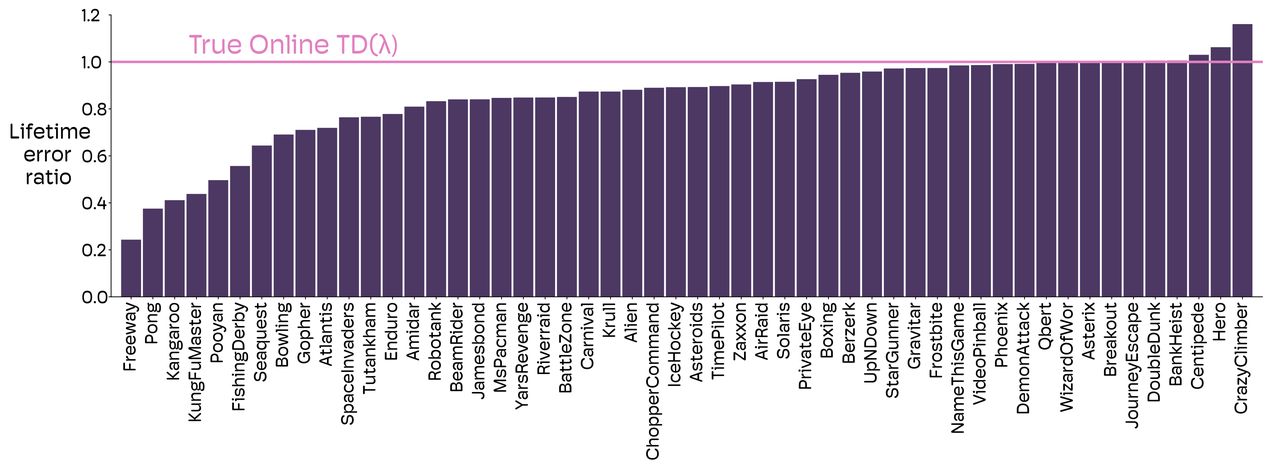
1) SwiftTD learned faster compared to the baseline



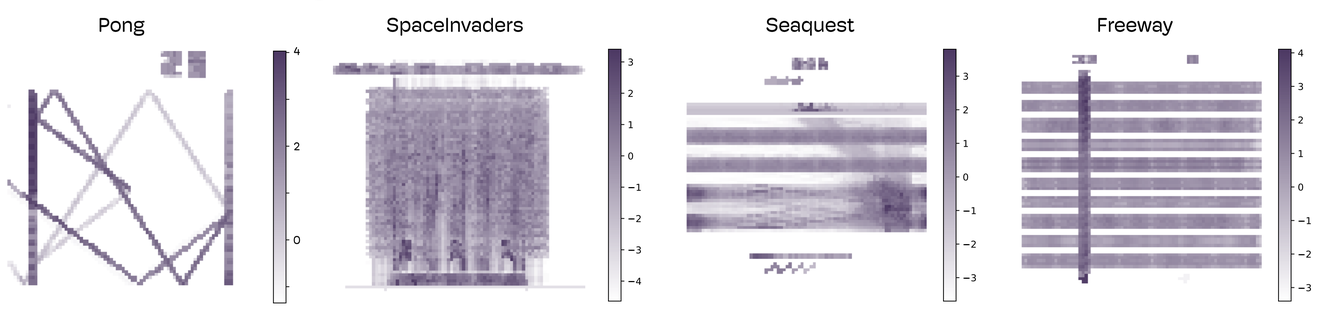
2) SwiftTD benefited from each of the three ideas



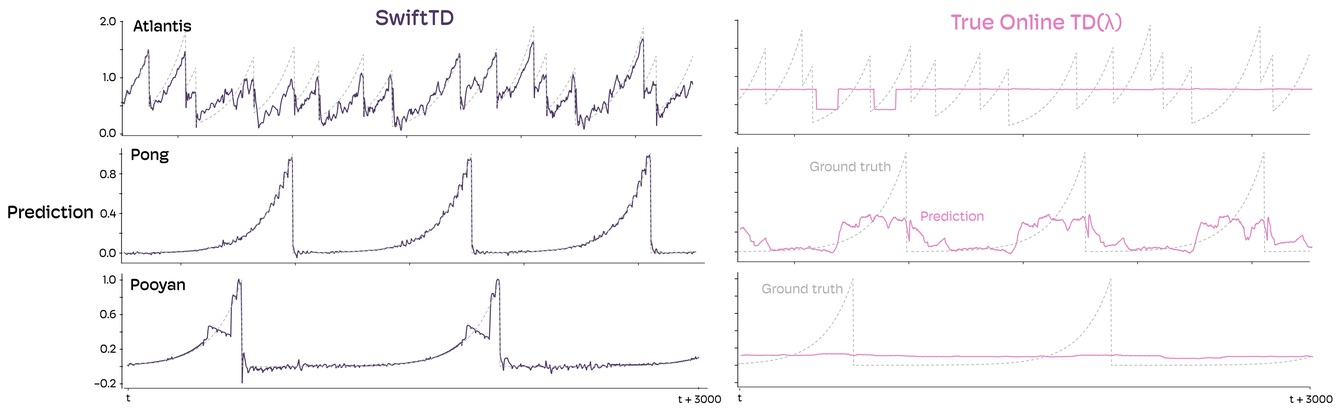
3) SwiftTD improved performance on the majority of the games



4) SwiftTD assigned credit to the relevant features



5) SwiftTD made more accurate predictions in final 3,000 steps



References

[1] Degris, T., Javed, K., Sharifnassab, A., Liu, Y., & Sutton, R. (2024). Step-size Optimization for Continual Learning. arXiv preprint arXiv:2401.17401..

[2] Javed, K., Shah, H., Sutton, R. S., & White, M. (2023). Scalable Real-time Recurrent Learning using Columnar-constructive Networks. The Journal of Machine Learning Research, 24(1), 12024-12057.

www.khurramjaved.com

mail@khurramjaved.com