**Prioritized Experience Replay**

Schaul et. al.

Carefully read the paper (pages 1-8 through discussion). We will discuss the questions below.

---

Questions for Discussion:

1) What was novel about this paper?

The novel approach presented in this paper was the idea of prioritizing the experiences (that is replaying more important experience more often) that are replayed, versus just sampling from a replay buffer uniformly.

2) What new feature does this approach bring to experience replay? How does it work?

It introduces the concept of learning potential by using the magnitude of the temporal difference (TD) error.

3) What is the issue of greedy error prioritization (Sec 3.3)? How is it overcome?

The agents would repeatedly see the same few “high error” transition. Since the same samples dominate, this can lead to overfitting or instability

It is also sensitive to noise spikes. A large TD error doesn’t always mean a transition in truly informative. Greedy prioritization would overemphasize noisy transitions.

To overcome these issues, they used stochastic sampling, which means they bias sampling toward high error transitions, but don’t exclude others completely. Low-error transitions still have a small chance of being replayed, maintaining diversity

 **Stochastic sampling** changes how often transitions are chosen.

 **Importance sampling** changes how much each chosen transition contributes.

**Greedy prioritization** focuses only on the largest TD-errors, which leads to overfitting, loss of diversity, and instability.  
**Prioritized Experience Replay** fixes this by using **stochastic prioritization**, where sampling is biased by TD-error but still includes all experiences with some probability.

4) Why is importance sampling used in the approach?

In PER high priority samples dominate, leading to bias. Importance sampling is used to correct this bias caused by non-uniform sampling.

| **Feature / Concept** | **Deep Q-Network (DQN)** | **Double DQN** | **Prioritized Experience Replay (PER)** |
| --- | --- | --- | --- |
| **Main Goal** | Handle **large/continuous state spaces** using a neural network as a function approximator. | **Reduce overestimation bias** in Q-values caused by the max operator. | **Improve sample efficiency** by replaying more informative transitions more often. |
| **Key Problem Addressed** | Correlated updates & instability in online Q-learning. | Q-values become **systematically too high**, harming policy quality. | Uniform replay wastes computation on uninformative samples. |
| **Core Innovation** | Adds **Experience Replay** (uniform) + **Target Network** to stabilize learning. | **Decouples action selection and evaluation** using two networks. | Introduces **non-uniform (priority-based) sampling** of transitions. |
| **Target Formula** | (Y^{DQN}=r+\gamma\max\_a Q(s',a;\theta^-)) | (Y^{DoubleDQN}=r+\gamma Q(s',\arg\max\_a Q(s',a;\theta);\theta^-)) | Same as DQN/Double DQN (used jointly), but sampling of transitions is prioritized by TD-error. |
| **Network Roles** | One **online** (θ) + one **target** (θ⁻) copied every τ steps. | Same two networks; online chooses action, target evaluates it. | Same base algorithm (DQN or Double DQN); modifies the **replay buffer** behavior. |
| **Sampling Rule** | Uniform random from buffer. | Uniform random. | (P(i)=\frac{( |
| **Bias Fix** | Uses a target network to stabilize but not bias-correct. | Removes **selection-evaluation coupling**, reducing overoptimism. | Adds **importance-sampling weights** (w\_i=(\tfrac{1}{N P(i)})^{\beta}) to correct sampling bias. |
| **Hyperparameters** | Replay size, learning rate, τ (target update freq). | Same + optional tweak to τ. | α (prioritization strength), β (IS correction), ε (small offset). |
| **Typical Outcome** | Learns good policies but sometimes unstable or overoptimistic. | More stable learning and better final performance. | Faster learning and higher sample efficiency. |
| **Main Limitation** | Overestimation bias; uniform replay wastes updates. | Requires two forward passes; still uses uniform replay. | Needs bookkeeping for priorities (sum-tree) and tuning α, β. |
| **Complexity Added** | Moderate (target network + buffer). | Minimal (new target formula). | Moderate (priority tree + weights). |
| **Empirical Gains (Atari)** | Baseline. | Higher median score (93 → 115 %). | Further improvement when combined with Double DQN. |