1 Main ideas

1. The paper proposes to more frequently replay transitions with high expected learning progress, as measured by the magnitude of their TD error.
2. They argue that such prioritization leads to loss of diversity, which they alleviate with stochastic prioritization.
3. And introduce bias, which they correct for with IS.

2 Prioritized Replay

2.1 Motivating example

1. They design a toy experiment where the replay buffer contains mostly uninformative transitions and few useful transitions.
2. There are 3 replay schemes tested: oracle, uniform, and 'greedy TD error prioritization'.
3. The oracle greedily selects the transition that maximally reduces the global loss in its current state (in hindsight, after the parameter update).
4. The 'greedy TD error prioritization' stores the last encountered TD along with each transition in the replay memory. The transition with the largest absolute TD error is replayed from the memory. New transition arrives arrive without a known TD-error, so we put them at maximal priority in order to guarantee that all experience is seen at least once.
5. The replay memory is filled by exhaustively executing all possible sequences of actions until termination in random order. This guarantee that exactly one sequence will succeed and hit the final reward, and all others will fail with zero reward.

2.2 Stochastic Prioritization

1. They argue that TD-error prioritization leads to 3 issues: transitions with low TD-error will not be replayed, transitions with high TD-error will be replayed frequently, leading to overfitting, and sensitive to noise spike.
2. To overcome these issues, they introduce a stochastic sampling method that interpolates between pure greedy prioritization and uniform random sampling.
3. This method ensures that the probability of being sampled is monotonic in a transition's priority, while guaranteeing a non-zero probability even for the transition with the lowest priority.
4. Concretely, the probability is: $P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$.
5. p_i is the priority of transition i.
6. The exponent α corresponds to how much prioritization is used, with $\alpha = 0$ corresponding to the uniform case.
7. They consider 2 variants: proportional and rank-based. In proportional, $p_i = |\delta_i| + \epsilon$ and in rank-based, $p_i = \frac{1}{rank(i)}$.

2.3 Annealing the bias

1. Instead of sampling uniformly from the buffer, TD-error prioritization and stochastic sampling changes the distribution of the stochastic updates.
2. To correct for this, they introduce an importance sampling weights

$$w_i = \left(\frac{1}{N} \cdot \frac{1}{P(i)}\right)^\beta$$

3. For stability reason, they always scale the weights by $1/max_i w_i$ so that they only scale the update downwards.
4. They anneal β so that it only reaches 1 at the end of learning.

3 Important details in the experiment section

1. Only a single hyperparameter adjustment was necessary compared to the baseline: Given that prioritized replay picks high-error transitions more often, the typical gradient magnitudes are larger, so we reduced the step-size η by a factor 4 compared to DQN.

4 Questions

1. Is there an extensive ablation study on the benefit of the importance sampling correction term? Figure 12 indicates that the IS term might not be even that important.
2. There does not seem to be an ablation studies on whether the issues stochastic sampling is supposed to mitigate exists and how much performance the stochastic sampling scheme introduces.
3. The paper mentions in Appendix A that the norm of the weight changed can also be used as the priority measure. But there does not seem to be any discussion as to what its performance is.