Prioritized Experience Replay

TD error delta

- Criteria used to assign priorities to each tuple
- Formula:
  $\delta_t = R_{t+1} + \gamma max_a \hat{q}(S_{t+1}, a, w) - \hat{q}(S_t, A_t, w)$δt =Rt+1+γmaxaq ^(St+1,a,w)−q ^(St,At,w)
- The bigger the error, the more we expect to learn from that tuple

Measure of Priority

- The magnitude of TD error
- Formula: $p_t = |\delta_t|$pt=|δt|
- Priority is stored along with each corresponding tuple in the replay buffer

Sampling Probability

- Computed from priority when creating batches
- Formula: $P(i) = \frac{p_i}{\sum_k p_k}$P(i)=∑kpkpi

Improvement on Prioritized Experience Relay

TD Error is Zero

- **Problem**: If the TD error is zero, then the priority

value of the tuple and hence its probability of being picked will also be zero. This doesn't necessarily mean we have nothing more to learn from such a tuple. It might be the case that our estimate was closed due to the limited samples we visited till that point.

- **Solution**: To prevent tuples from being starved for selection, we can add a small constant ee to every priority value. Thus, **priority** will be expressed as

$p_t = |\delta_t| + e$pt=|δt|+e

Greedy Usage of Priority Values

- **Problem**: Greedily using priority values may lead to a small subset of experiences being relayed over and over, resulting in a overfitting to that subset.
- **Solution**: Reintroduce some element of uniform random sampling. This adds another hyperparameter aa which we use to redefine the sample probability as

$P\left(i\right) = \frac{{p_i}^a}{\sum_k {p_k}^a}$P(i)=∑kpkapia

Adjustment to the Update Rule

We we use prioritized experience relay, we have to make one adjustment to our update rule, which is

$\Delta w = \alpha\left(\frac{1}{N} \cdot \frac{1}{Pi}\right)^b \delta_i \nabla_w \hat{q}\left(S_i, A_i, w\right)$Δw=α(N1·Pi1)bδi∇w q^(Si,Ai,w)

where $\left(\frac{1}{N} \cdot \frac{1}{Pi}\right)^{b}$(N1·Pi1)b stands for the importance-sampling weight.