经验回放让在线强化学习代理能够记住并重用过去的经验。 在之前的工作中，经验转换是从重放记忆中统一采样的。 然而，这种方法只是以最初经历的相同频率重播经验，而不管它们的重要性。在本文中，我们开发了一个优先考虑经验的框架，以便更频繁地重播重要的经验，从而更有效地学习。 我们在深度 Q 网络 (DQN) 中使用优先经验回放机制，DQN是一种强化学习算法，在许多 Atari 游戏中实现了人类水平的性能。 具有优先经验回放的 DQN 达到了新的先进水平，在49场比赛中，有41场比赛效果优于基于传统回放的DQN。

在线强化学习 (RL) 智能体在观察经验流时会逐步更新其参数（策略、价值函数或模型）。 在最简单的模型中，它们会在一次更新后立即丢弃传入的数据。 基于此存在两个问题：

* 基于梯度的更新打破了数据之间的独立同分布(i.i.d.) 假设问题
* 快速忘记了一些罕见重要的、以后使用的经验数据

经验回放(Lin, 1992)解决了这两个问题：将经验存储在回放内存中，通过混合最近和过去经验进行更新，可以打破时间上的相关性，并且使罕见经验将不仅仅用于一次更新。这在深度 Q 网络 (DQN) 算法 (Mnih et al., 2013; 2015) 中得到了证明，该算法通过使用经验回放来稳定由深度神经网络表示的价值函数的训练。 具体来说，DQN 使用了一个大的滑动窗口重放记忆，从其中随机均匀采样，并且平均重新访问每个经验八次。 一般来说，经验回放可以减少学习所需的经验量，并用更多的计算和更多的内存来代替它——这些资源通常比RL智能体与环境的交互更易得。

在本文中，我们研究了如何重放经验可以使其比统一随机均匀重放所有经验更高效。关键思想是RL智能体可以从某些经验比其他经验中学到更多。一些经验可能或多或有效、多余或与任务相关。 一些经验可能不会立即对智能体有用，但当智能体能力增加时可能会变得有用（Schmidhuber，1991）。经验重放将在线学习智能体从按照他们经验的固有顺序进行处理中解放出来。优先经验重放进一步将智能体从以相同概率重放经验的过程中解放出来。

特别是，我们建议更频繁地重放具有高预期学习效率的经验，通过它们的时间差 (TD) 误差的大小来衡量。这种优先级可能导致多样性的损失，我们利用随机优先级采样、偏置和重要性采样来避免该问题。我们得到的算法是稳健且可扩展的，我们在 Atari 2600上进行了演示，在环境中我们获得了更快的学习和先进的性能。

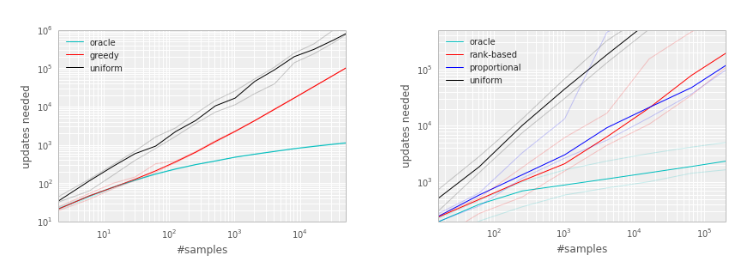
**优先经验回放**

使用重放经验需要在两个层次上进行设计选择：存储哪些经验，以及重放哪些经验（以及如何进行）。 本文只讨论后者：最有效地利用回放记忆进行学习，它存储哪些经验不在我们的讨论范围内。

**使用时序差分误差进行优先级排序**

优先回放的核心部分是衡量每个经验重要性的标准。一个理想化的标准是RL智能体可以从其当前状态的经验中学习的量（预期的学习进度）。虽然这个度量不能直接访问，但可以合理的使用TD 误差 δ 的大小代替，它表明经验是多么有用：具体来说，该值是其与下一步估计值相差多远（Andre 等人，1998 年）。这特别适用于增量的在线 RL 算法，例如 SARSA 或 Q-learning，它们已经计算了 TD 误差并按照 δ 的比例更新参数。 在某些情况下，TD 误差也可能是一个糟糕的估计，例如当奖励嘈杂时； 有关替代方案的讨论，请参见附录 A

为了证明通过 TD 误差优先经验重放的有效性，我们在 Blind Cliffwalk 将基于“贪婪的 TD 错误优先级”uniform and oracle baselines算法与进行比较。该算法将最后遇到的TD误差与每个经验一起存储在重放存储器中。从存储器中重放具有最大绝对 TD 误差的经验。Q\_learning基于此经验更新，它根据 TD 误差按比例更新权重。新的经验刚获得时并不知道TD误差值，因此我们将它们置于最高优先级，以保证所有经验至少被使用一次。 图 2（左）显示，该算法大大减少了解决盲崖行走任务所需的工作量。



**随机优先级**

然而，贪婪的 TD 错误优先级有几个问题。首先，为了避免对整个重放经验进行的扫描，TD 错误仅针对重放经验进行更新。 一个后果是第一次访问时 TD 误差较低的经验可能很长时间不会得到重放（这意味着可能永远不会使用滑动窗口重放该经验）。此外，它对噪声峰值很敏感（例如，当奖励是随机的时），这可能会因自举而加剧，另外近似误差是另一个噪声源。最后，贪心优先级专注于一小部分经验：错误缩小缓慢，尤其是在使用函数逼近时，这意味着最初的高错误经验会频繁重播。这种缺乏多样性使得系统容易过度拟合。

为了克服这些问题，我们引入了一种随机抽样方法，该方法结合纯贪婪优先级和均匀随机抽样。我们确保在经验的优先级中被采样的概率是单调的，同时即使对于最低优先级的经验也保证非零概率。具体来说，我们将采样转换 i 的概率定义为

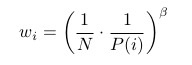


其中pi为第i个经验的priority， 用于调节优先程度（α=0的时候退化为均匀采样)。

我们考虑的第一个变体是直接的比例优先级，其中 pi = |δi|+ε，其中ε是一个小的正常数，它可以防止过渡的边缘情况在经验的TD误差为0时不被访问。 第二种变体是间接的、基于等级的优先级， ,当重放经验根据 |δi| 储存时，rank(i)是经验i的等级。 在这种情况下，P变为指数为 α 的幂律分布。 两种分布在 |δ| 中都是单调的，但后者可能更稳健，因为它对异常值不敏感。 如图 2（右）所示，随机优先级的两种变体都会导致 Cliffwalk 任务的相比与统一基线大幅加速。

**退火偏置**

随机更新对期望值的估计依赖于与预期相同的分布对应的那些更新。 优先重放会引入偏差，因为它以一种不受控制的方式改变了这种分布，因此改变了估计将收敛到的解决方案（即使策略和状态分布是固定的）。 我们可以通过使用重要性采样 (IS) 权重来纠正这种偏差

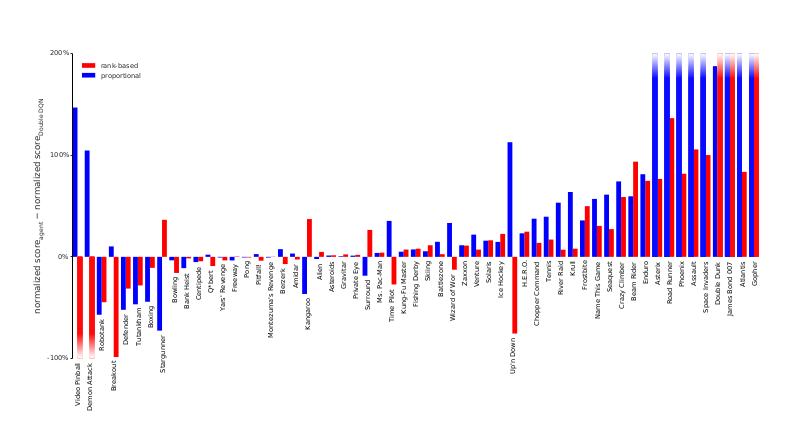


如果 β = 1，则完全补偿非均匀概率 P (i)。这些权重可以通过使用 ωiδi 而不是δi折叠到 Q 学习更新中（因此这是加权 IS，而不是普通 IS，请参阅 例如 Mahmood 等人，2014年）。出于稳定性原因，我们总是将权重归一化 1/ maxi wi，这样它们只会向下缩放更新。

在典型的强化学习场景中，更新的无偏性在训练结束时的收敛附近最为重要，因为由于策略、状态分布和引导目标的变化，该过程无论如何都是高度非平稳的； 我们假设在这种情况下可以忽略一个小的偏差（另请参见附录中的图 12，了解 Atari 上的完整 IS 校正的案例研究）。 因此，我们通过在指数 β 上定义一个仅在学习结束时达到 1 的时间表，来利用随时间退火重要性采样校正量的灵活性。 在实践中，我们将 β 从其初始值 β 0 线性退火到 1。请注意，此超参数的选择与优先级指数 α 的选择相互作用； 同时增加两者会更积极地优先采样，同时更强烈地对其进行校正。

在非线性函数逼近（例如深度神经网络）的情况下，重要性采样与优先重放相结合时还有另一个好处：这里的大步可能非常具有破坏性，因为梯度的一阶逼近仅在局部可靠，并且具有 用较小的全局步长来防止。 相反，在我们的方法中，优先级可确保多次看到高误差转换，而 IS 校正减少了梯度幅度（从而减少了参数空间中的有效步长），并允许算法遵循高度非线性优化的曲率 景观，因为泰勒展开不断地重新近似。

基于最先进的 Double DQN 算法，我们将我们的优先重放算法组合成一个全面的强化学习代理。 我们的主要修改是将 Double DQN 使用的均匀随机抽样替换为我们的随机优先级和重要性抽样方法（参见算法 1）。



# PPO

我们提出了一系列用于强化学习的新策略梯度方法，通过与环境的交互对数据进行采样，并使用随机梯度上升优化“智能体”目标函数。 虽然标准策略梯度方法对每个数据样本只执行一次梯度更新，但我们提出了一种新的目标函数，可以实现小批量更新多个时期。 我们称之为邻近策略优化 (PPO) 的新方法具有信任区域策略优化 (TRPO) 的一些好处，但它实现起来更简单、更通用，并且具有更好的样本复杂度（根据经验）。 我们的实验在一组基准任务上测试了 PPO，包括模拟机器人运动和 Atari 游戏，我们表明 PPO 优于其他在线策略梯度方法，并且总体上在样本复杂性、简单性和执行时间之间取得了良好的平衡。