强化学习中的随机过程

https://blog.csdn.net/shengzimao/article/details/130939571

在应用和构建强化学习过程的时候这些定理的条件需要满足并给出证明。例如Soft Actor-Critic的论文中由于在奖励中加入了信息熵，并且在强化学习过程中使用了软最大值，所以从贝尔曼方程到随机估计过程的构建都需要重新审查，并给出严格的数学证明。没有这些严格的数学证明，提出的新方法就不能保证有效且被研究社区接受。还有一个重要的原因是，当我们在工程实践中，在数学建模的过程完成后，如果这些定理都满足了，那么我们的实践便有了可靠的支撑，强化学习的算法很容易在应用中出现效果不好的情况，此时如果有了这些定理的支撑，便可以坚定的去审查代码实现或者数据是否有问题，不至于使自己淹没在巨大的不确定的潜在问题中。

收缩映射定理

定理内容

收缩映射定理（Contraction mapping theorem） 对于任意的形如 $x = f(x)$ 的等式，其中 x 和 $f(x)$ 都是实向量，如果 $f(\cdot)$ 是一个收缩映射，那么将有如下性质：

1. 存在性：存在一个固定点 x^* 满足 $f\left(x^*\right) = x^*$
2. 唯一性：固定点 x^* 是唯一的
3. 算法：考虑迭代过程

$$x_{k+1} = f\left(x_k\right)$$

其中 $k = 0, 1, 2, \ldots$。当 $k \to \infty$ 的时候， $x_k \to x^*$ ，并且收敛的速度是指数级的。

贝尔曼最优方程的简洁矩阵形式为

$$v = \max_{\pi \in \Pi}\left(r_\pi + \gamma P_\pi v\right)$$

其中 $v \in \mathbb{R}^{|\mathcal{S}|}$ 且 $\max\limits_{\pi}$ 是在每个元素上操作的， r_π 和 P_π 的结构和常规的贝尔曼方程相同：

$$[r_\pi]_s = \sum_{a \in \mathcal{A}} \pi\left(a \mid s\right) \sum_{r \in \mathcal{R}} p\left(r \mid s, a\right) r, \qquad [P_\pi]_{s,s'} = p\left(s' \mid s\right) \doteq \sum_{a \in \mathcal{A}} \pi\left(a \mid s\right) p\left(s' \mid s, a\right)$$

再定义 $f(v) \doteq \max\limits_{\pi \in \Pi}\left(r_\pi + \gamma P_\pi v\right)$ ，于是贝尔曼最优方程可以表达为

$$v = f(v)$$

贝尔曼最优方程的收缩映射性质 上述贝尔曼最优方程右边的函数 $f\left(v\right)$ 是一个收缩映射。特别的，对于任意的 $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$ 有以下不等式成立

$$\left\|f\left(v_1\right) - f\left(v_2\right)\right\|_\infty \leqslant \gamma \left\|v_1 - v_2\right\|_\infty$$

其中 $\gamma \in (0,1)$ 为折扣率，$\|\cdot\|_\infty$ 是最大范数，表示一个向量所有元素的绝对值的最大值。同时满足上述收缩映射定理的存在性与唯一性以及求解算法。

证明

TODO

Dvoretzky’ s 收敛定理

定理内容

考虑一个随机过程 $\Delta_{k+1} = (1-\alpha_k)\Delta_k + \beta_k \eta_k$，其中 $\{\alpha_k\}_{k=1}^{\infty}$，$\{\beta_k\}_{k=1}^{\infty}$，$\{\eta_k\}_{k=1}^{\infty}$ 是随机序列。当满足以下条件:

1. $\sum_{k=1}^{\infty} \alpha_k = \infty, \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \sum_{k=1}^{\infty} \beta_k^2 < \infty \,\text{uniformly almost surely}$；
2. $\mathbb{E}[\eta_k | H_k] = 0 \,\text{and}\, \mathbb{E}[\eta_k^2 | H_k] \leqslant C \,\text{almost surely}$；

则 Δ_k 会几乎必然（almost surely）收敛到 0。

证明

假设 α_k, β_k 可以由 H_k 完全确定，即 $\alpha_k = \alpha_k(H_k)$，$\beta_k = \beta_k(H_k)$。则:

$$\mathbb{E}[\alpha_k | H_k] = \alpha_k, \quad \mathbb{E}[\beta_k | H_k] = \beta_k$$

构造 $h_k = \Delta_k^2$，可得

$$\begin{aligned}
\mathbb{E}[h_{k+1} - h_k | H_k] &= \mathbb{E}\left[\Delta_{k+1}^2 - \Delta_k^2 | H_k\right] \\
&= \mathbb{E}\left[-\alpha_k(2-\alpha_k)\Delta_k^2 + \beta_k^2 \eta_k^2 + (2-2\alpha_k)\beta_k \eta_k \Delta_k | H_k\right] \\
&= -\alpha_k(2-\alpha_k)\Delta_k^2 + \beta_k^2 \mathbb{E}\left[\eta_k^2 | H_k\right] + (2-2\alpha_k)\beta_k
\end{aligned}$$

因为 $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$，所以 $\alpha_k \to 0$。

所以存在 N，当 $k > N$ 时有 $\alpha_k \leqslant |\alpha_k| < 1$ (极限定义)，$-\alpha_k(2-\alpha_k)\Delta_k^2 < 0$，此时 $\mathbb{E}\left[h_{k+1} - h_k | H_k\right] \leqslant \beta_k^2 C, k > N$。又因为 $\sum_{k=1}^{\infty} \beta_k^2 = C_{\beta^2} < \infty$，因此有:

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{E}[h_{k+1} - h_k | H_k] &= \left(\sum_{k=1}^{N} + \sum_{k=N+1}^{\infty}\right) \mathbb{E}\left[h_{k+1} - h_k \mid H_k\right] \\
&\leq \sum_{k=1}^{N} \mathbb{E}\left[h_{k+1} - h_k \mid H_k\right] + \sum_{k=N+1}^{\infty} \beta_k^2 C \\
&\leq \sum_{k=1}^{N} \mathbb{E}\left[h_{k+1} - h_k \mid H_k\right] + C_{\beta^2} C < \infty
\end{aligned}$$

又因为

$$\begin{aligned}\sum_{k=1}^{\infty}\alpha_k\Delta_k^2=\sum_{k=N}^{\infty}\alpha_k\Delta_k^2+\sum_{k=1}^{N}\alpha_k\Delta_k^2&<\sum_{k=1}^{N}\alpha_k\Delta_k^2+\sum_{k=N}^{\infty}\alpha_k\left(2-\alpha_k\right)\Delta_k^2\\&<\sum_{k=1}^{N}\alpha_k\Delta_k^2+\sum_{k=1}^{\infty}\alpha_k\left(2-\alpha_k\right)\Delta_k^2\\&<\sum_{k=1}^{N}\alpha_k\Delta_k^2-\sum_{k=1}^{\infty}\mathbf{E}\left[h_{k+1}-h_k\mid H_k\right]+\sum_{k=1}^{\infty}\beta_k^2C\end{aligned}$$

由之前的证明的 $\mathbb{E}[h_{k+1}-h_k|H_k]<\infty$，$\sum_{k=1}^{\infty}\beta_k^2C=C_{\beta^2}C<\infty$ 可知：

$$\sum_{k=1}^{\infty}\alpha_k\Delta_k^2<\infty$$

因此几乎必然有 $\Delta_k\to0$，证明完毕。

Robbins-Monro定理

定理内容

对于求解 $g(w)=0$ 的迭代算法

$$w_{k+1}=w_k-a_k\tilde{g}\left(w_k,\eta_k\right)\qquad k=1,2,3,\ldots$$

其中 w_k 是对方程根的第 k 个估计，$\tilde{g}\left(w_k,\eta_k\right)$ 是第 k 个带噪声的观测，α_k 是一个正系数。如果下列条件

1. $0<c_1\leq\nabla_w g(w)\leq c_2$ for all w
2. $\sum_{k=1}^{\infty}a_k=\infty\quad\text{and}\quad\sum_{k=1}^{\infty}=a_k^2<\infty$
3. $\mathbb{E}[\eta_k|H_k]=0\quad\text{and}\quad\mathbb{E}\left[\eta_k^2|H_k\right]<\infty$

其中 $\mathcal{H}=\{w_k,w_{k-1},\ldots\}$，那么 w_k 几乎必然收敛到根 w^* 满足 $g\left(w^*\right)$。该算法称为Robbins-Monro算法。

证明

上述Robbins-Monro算法可以进一步写成

$$\begin{aligned}w_{k+1}&=w_k-a_k\tilde{g}\left(w_k,\eta_k\right)\\&=w_k-a_k\left[g\left(w_k\right)-\eta_k\right]\end{aligned}$$

从而

$$w_{k+1}-w^*=w_k-w^*-a_k\left[g\left(w_k\right)-g\left(w^*\right)-\eta_k\right]$$

又根据均值定理，我们有

$$g\left(w_k\right)-g\left(w^*\right)=\nabla_w g\left(w_k^{\prime}\right)\left(w_k, w^*\right)$$

其中 $w_k^{\prime} \in\left[w_k, w^*\right]$。令 $\Delta_k \doteq w_k-w^*$。上面的等式变成

$$\begin{aligned}\Delta_{k+1} &= \Delta_k-a_k\left[\nabla_w g(w_k^{\prime})(w_k-w^*)+\eta_k\right] \\ &= \Delta_k-a_k \nabla_w g(w_k^{\prime}) \Delta_k+a_k(-\eta_k) \\ &= [1-\underbrace{a_k \nabla_w g(w_k^{\prime})}_{\alpha_k}] \Delta_k+a_k(-\eta_k)\end{aligned}$$

注意根据定理假设条件 $\nabla_w g(w)$ 是有上下界的，满足 $0<c_1 \leqslant \nabla_w g(w) \leqslant c_2$ ，又因为有定理假设条件2，那么Dvoretzky定理的所有的条件都满足，因此 $\lim _{k \rightarrow \infty} \Delta_k=0$ a.s. 。

应用

Robbins-Monro定理可以用来估计数学期望，这在强化学习中经常用到。可以构造 $g(w) \doteq w-\mathbb{E}[X]$ ，此时问题变成求解方程 $g(w)=0$。给定一个值 w ，带噪声的观察为 $\tilde{g} \doteq w-x$ ，其中 x 是对 X 的采样。 $\tilde{g}$ 可以写成

$$\begin{aligned}\tilde{g}(w, \eta) &= w-x \\ &= w-x+\mathbb{E}[X]-\mathbb{E}[X] \\ &= (w-\mathbb{E}[X])+(\mathbb{E}[X]-x) \doteq g(w)+\eta\end{aligned}$$

其中 $\eta \doteq \mathbb{E}[X]-x$。此时估算期望的Robbins Monro算法为

$$w_{k+1}=w_k-\alpha_k \tilde{g}\left(w_k, \eta_k\right)$$

拓展Dvoretzky’ s 收敛定理

拓展的Dvorerzky's定理是一个可以处理多变量的更通用的定理，可以被用来分析随机迭代算法的收敛性，在Q-learning中会有应用。

定理内容

考虑一个实数有限集合 $\mathcal{S}$ ，对于随机过程

$$\Delta_{k+1}(s)=(1-\alpha_k(s)) \Delta_k(s)+\beta_k(s) \eta_k(s)$$

对于每一个 $s \in \mathcal{S}$ ， $\Delta_k(s)$ 将几乎必然收敛到 0 只要以下条件满足：

1. $\sum_{k=1}^{\infty} \alpha_k(s)=\infty, \sum_{k=1}^{\infty} \alpha_k^2(s)<\infty, \sum_{k=1}^{\infty} \beta_k^2(s)<\infty$
   $and \quad \mathbb{E}\left[\beta_k^2(s) \mid H_k\right] \leqslant \mathbb{E}\left[\alpha_k^2(s) \mid H_k\right]$ uniformly almost surely
2. $\left\|\mathbb{E}\left[\eta_k^2 \mid H_k\right]\right\|_{\infty} \leqslant \gamma\left\|\Delta_k\right\|_{\infty}$ ， 其中$\gamma \in(0,1)$
3. $\operatorname{var}\left[\eta_k(s) \mid \mathcal{H}_k\right] \leqslant C\left(1+\left\|\Delta_k(s)\right\|_{\infty}\right)^2$ ， 其中C为一个常数

这里 $\mathcal{H}_k = \{\Delta_k, \Delta_{k-1}, \ldots, \eta_{k-1}, \ldots, \alpha_{k-1}, \ldots, \beta_{k-1}, \ldots\}$ 代表历史信息，$\|\cdot\|_\infty$ 是最大范数，表示一个向量所有元素的绝对值的最大值。

有以下要点需要注意

- 拓展的Dvorerzky's定理可以处理多个状态的强化学习问题。
- 在应用定理的时候，我们需要明确指出定理条件在集合 $\mathcal{S}$ 中的每一个状态（或者状态动作对）上都成立。

证明

TODO

随机梯度下降（SGD）

随机梯度下降法的核心思想是用随机的梯度代替真实的梯度。

定理内容

对于最优化问题 $\min_w J(w) = \mathbb{E}\left[f\left(w, X\right)\right]$，当使用随机过程迭代式 $w_{k+1} = w_k - \alpha_k \nabla_w f\left(w_k, x_k\right)$ 进行参数迭代时，若满足

$$
\begin{aligned}
&(1) \quad 0 < c_1 \leq \nabla_w^2 f(w, X) \leq c_2; \\
&(2) \quad \sum_{k=1}^{\infty} \alpha_k = \infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty; \\
&(3) \quad \{x_k\}_{k=1}^{\infty} \ \text{ i.i.d}
\end{aligned}
$$

则 $w_k \to w^*$，其中 $\nabla_w \mathbb{E}\left[f\left(w^*, X\right)\right] = 0$ a.s.

证明

TODO

马尔可夫过程的稳态分布定理

稳态分布定理的核心含义是，在满足条件的马尔可夫链中，随着时间的推移，链的状态分布会逐渐趋向于一个固定的分布，且这个分布与初始状态无关。有以下要点：

1. 时间无关的长期行为

   无论马尔可夫链的初始状态是什么，经过足够长的时间，系统将不再依赖于起始状态，且进入长期的平稳状态。换句话说，系统最终会在各个状态之间按照稳态分布频率进行转移。

2. 不可约性和遍历性

   定理要求马尔可夫链是**不可约**的，即从任意一个状态可以通过有限步到达任意其他状态（状态之间可以相互连通）。还要求链是**正则**或**遍历的**，即所有状态都可以在有限时间内返回自身。这两个性

质保证了系统不会停留在某个特定状态或某个状态集合中。

3. 稳态分布的物理解释

   在很多实际应用中，稳态分布代表了系统在长期运行中的各状态的“停留频率”或“访问概率”。例如：

   - 在排队论中，稳态分布可以描述在长期运行的服务系统中，系统处于不同状态的概率。
   - 在强化学习中，稳态分布可以表示在马尔可夫决策过程（MDP）中，某一策略下的状态分布，这对于理解策略的长期效果很重要。

4. 计算稳态分布

   稳态分布 π 可以通过求解方程组 $\pi P = \pi$ 和 $\sum_i \pi_i = 1$ 来计算。对于大规模的状态空间，常用数值方法来近似求解，比如迭代法或幂法（Power Iteration）。

在强化学习的值函数估计的方法中，正是因为稳态分布的存在，才可以将优化目标中的随机变量 S 写成 s_t ，从而可以使用随机梯度的方法。

定理内容

设Markov Process的状态空间为 $\mathcal{S}$ ，状态量的个数为 $\mathcal{S}$ ，定义在策略 π 下的状态转移概率矩阵 $P_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ ，定义 k 步状态转移矩阵 $P_\pi^k = \left\{p_{i,j}^{(k)}\right\}_{|\mathcal{S}|\times|\mathcal{S}|}$

$$p_{ij}^{(k)}(\pi) = \mathbf{Prob}(S_{t_k} = j \mid S_{t_0} = i, \pi)$$

其满足 $P_\pi^k = P_\pi P_\pi^{k-1}$ 。对于任意一个初始状态分布 $d_0 \in \mathbb{R}^{|\mathcal{S}|}$ ，在策略 π 下经过 k 轮迭代后的状态分布为 d_k ，且有 $d_k^\top = d_0^\top P_\pi^k$ 。

若对于任意的两个状态 $s_i, s_j \in \mathcal{S}$ ，都存在有限的步长 k ，使得 $\left[P_\pi^k\right]_{ij} > 0$ ，则有如下结论：

(1) $P_\pi^k \to \frac{1}{|S|} d_\pi^\top;$

(2) $d_\pi^k \to d_\pi^0 \frac{1}{|S|} d_\pi^\top = d_\pi^T;$

(3) $d_\pi^\top$ 满足 $d_\pi^\top = d_\pi^\top P.$

此时称这样的马尔可夫过程是正则的。