ゼロから作る Deep Learning ④のノート

2026 年 2 月 15 日

概要

手書きの方が速いと感じたので、ここでのノートは一旦中止にする（再開するかもしれない）。

1　ベルマン方程式

1.1　ベルマン方程式の導出

まずは、時刻の取り扱いについて定義する。

定義 1.1. 時刻 t における状態を S_t と表し、同時刻のその状態での行動を A_t と表す。またその行動によって得られた報酬を R_t と表すことにする。

定義 1.2. 次のような定義を与える。

- 時刻 t での状態 s $(S_t = s)$ から行動 a $(A_t = a)$ が選ばれる確率を**方策**といい、$\pi(s \mid a)$ と表す。
- ある状態 s から行動 a によって次の状態 s' に遷移する確率を**状態遷移確率**といい、$p(s' \mid s, a)$ と表す。
- 状態 s から行動 a をとり、次の状態 s' に遷移した際に得られる報酬を $r(s, a, s')$ と表す。
- 収益 G_t とは、時刻 t において、その後で得られる全ての報酬の価値を表したものである。

$$G_t = \sum_{k=t}^{\infty} \gamma^{k-t} R_k,\ \gamma \in [0, 1)$$

定義 1.3 (状態価値関数)**. 状態価値関数** $v_\pi(s)$ とは、方策を π とし、時刻 t における状態が s であるときに、その後の行動や遷移する状態のすべてを加味した際の収益の平均（期待値）のことで

ある。

すなわち、

$$v_\pi(s) = \mathbb{E}_\pi \left(G_t \mid S_t = s \right)$$

である。

時刻 t での状態価値関数 $v_\pi(s)$ からある行動によって遷移し、時刻 $t+1$ での状態価値関数を $v_\pi(s')$ とする。このとき、次のような方程式が成り立つ。