Book Errata

The following is the errata for the published book *Foundations of Reinforcement Learning with Applications in Finance* by *Ashwin Rao* and *Tikhon Jelvis*

Chapter 2

On page 35, the two definitions of `sqrt` need to initialize `x_n`:

```
def sqrt(a: float) -> float:
    x = a / 2 # initial guess
    x_n = a
    while abs(x_n - x) > 0.01:
        x = x_n
        x_n = (x + (a / x)) / 2
    return x_n
```

and

```
def sqrt(a: float, threshold: float) -> float:
    x = a / 2 # initial guess
    x_n = a
    while abs(x_n - x) > threshold:
        x = x_n
        x_n = (x + (a / x)) / 2
    return x_n
```

On page 37, the definition of `converge` needs one level less indentation:

```
def converge(values: Iterator[float], threshold: float) -> Iterator[float]:
    for a, b in itertools.pairwise(values):
        yield a

        if abs(a - b) < threshold:
            break
```

Chapter 3

Definition of Stationary Distribution of a Markov Process

On page 58, the equation in Definition 3.7.1 that defines the Stationary Distribution of a Markov Process is incorrect. The corrected equation is as follows:

$$\pi(s') = \sum_{s \in \mathcal{N}} \pi(s) \cdot \mathcal{P}(s, s') \text{ for all } s' \in \mathcal{N}$$

Definition of $\mathcal{R}_T$ for `SimpleInventoryMRPFinite`

On page 67, there is an error in the equation for $\mathcal{R}_T$ for the case when the next state's (S_{t+1}) On-Hand is equal to 0. The corrected equation is as follows:

$$\mathcal{R}_T((\alpha,\beta),(0,C-(\alpha+\beta))) = -h\alpha - p\frac{\sum_{j=\alpha+\beta+1}^{\infty} f(j)\cdot(j-(\alpha+\beta))}{\sum_{j=\alpha+\beta}^{\infty} f(j)}$$

$$= -h\alpha - p(\lambda - (\alpha+\beta)(1-\frac{f(\alpha+\beta)}{1-F(\alpha+\beta-1)}))$$

The corresponding correction to the code on page 68 involves changing the assignment to the variable `reward` according to the following code:

```
reward = base_reward - self.stockout_cost * \
    (self.poisson_lambda - ip *
     (1 - self.poisson_distr.pmf(ip) / probability))
```

The full body of this (corrected) code can be found in: rl/chapter2/simple_inventory_mrp.py.

Chapter 4

Definition of $\mathcal{R}_T$ for `SimpleInventoryMDPCap`

On page 86, there is an error in the equation for $\mathcal{R}_T$ for the case when the next state's (S_{t+1}) On-Hand is equal to 0. The corrected equation is as follows:

$$\mathcal{R}_T((\alpha,\beta),\theta,(0,\theta)) = -h\alpha - p\frac{\sum_{j=\alpha+\beta+1}^{\infty} f(j)\cdot(j-(\alpha+\beta))}{\sum_{j=\alpha+\beta}^{\infty} f(j)}$$

$$= -h\alpha - p(\lambda - (\alpha+\beta)(1-\frac{f(\alpha+\beta)}{1-F(\alpha+\beta-1)}))$$

The corresponding correction to the code on page 87 involves changing the assignment to the variable `reward` according to the following code:

```
reward: float = base_reward - self.stockout_cost * \
    (self.poisson_lambda - ip *
     (1 - self.poisson_distr.pmf(ip) / probability))
```

The full body of this (corrected) code can be found in: rl/chapter3/simple_inventory_mdp_cap.py.

Bellman Policy Equation Derivation

At the very first line of page 90, we say that Equation (4.3) can be obtained by combining Equation (4.1) and Equation (4.2). This is not correct. Rather, Equation (4.3) can be derived as follows:

Since $Q^\pi(s,a) = \mathbb{E}_{\pi,\mathcal{P}_R}[G_t|S_t=s,A_t=a]$, by expanding G_t as $R_{t+1}+\gamma\cdot R_{t+2}+\gamma^2\cdot R_{t+3}+\ldots$, we can expand $Q^\pi(s,a)$ as:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{P}_R}[R_{t+1}|S_t = s, A_t = a] + \gamma \cdot \mathbb{E}_{\pi,\mathcal{P}_R}[R_{t+2}|S_t = s, A_t = a] + \gamma^2 \cdot \mathbb{E}_{\pi,\mathcal{P}_R}[R_{t+3}|S_t = s, A_t = a] \\
&\quad + \ldots \\
&= \mathcal{R}(s,a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s,a,s') \sum_{a' \in \mathcal{A}} \pi(s',a') \cdot \mathcal{R}(s',a') \\
&\quad + \gamma^2 \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s,a',s') \sum_{a' \in \mathcal{A}} \pi(s',a') \sum_{s'' \in \mathcal{N}} \mathcal{P}(s',a'',s'') \sum_{a'' \in \mathcal{A}} \pi(s'',a'') \cdot \mathcal{R}(s'',a'') \\
&\quad + \ldots \\
&= \mathcal{R}(s,a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s,a,s') \cdot \mathcal{R}^\pi(s') + \gamma^2 \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s,a,s') \sum_{s'' \in \mathcal{N}} \mathcal{P}^\pi(s',s'') \cdot \mathcal{R}^\pi(s'') + \ldots \\
&= \mathcal{R}(s,a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s,a,s') \cdot (\mathcal{R}^\pi(s') + \gamma \cdot \sum_{s'' \in \mathcal{N}} \mathcal{P}^\pi(s',s'') \cdot \mathcal{R}^\pi(s'') + \ldots) \\
&= \mathcal{R}(s,a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s,a,s') \cdot V^\pi(s')
\end{aligned}
$$

The last line in the sequence of equations above is obtained by observing Equation (3.1) in Chapter 3, with the MRP Value Function expansion applied to the π-implied MRP for state s'.

Chapter 6

`argmax` code correction

On page 144, the body of the code for the method `argmax` (of `class FunctionApprox`) is not correct. The correct code is shown below and can also be found in rl/function_approximation.py.

```
    def argmax(self, xs: Iterable[X]) -> X:
        args: Sequence[X] = list(xs)
        return args[np.argmax(self.evaluate(args))]
```

Chapter 13

On page 374, in the second paragraph of Section 13.7.2, we say that $\boldsymbol{B}^\pi$ is a linear operator in vector space $\mathbb{R}^n$. This is an incorrect statement. $\boldsymbol{B}^\pi$ is an affine operator and not a linear operator because there is a vector-shift of $\mathcal{R}^\pi$ after multiplying $\boldsymbol{V}$ with the matrix $\gamma\mathcal{P}^\pi$. This second paragraph of Section 13.7.2 is corrected as follows:

Note that $\boldsymbol{B}^\pi$ is an affine transformation on vectors in $\mathbb{R}^n$. We lighten notation for application of the $\boldsymbol{B}^\pi$ operator on any vector $\boldsymbol{V} \in \mathbb{R}^n$ as simply $\boldsymbol{B}^\pi \cdot \boldsymbol{V}$ (with $\cdot$ conveying the notion of operator application). We've learnt that $\boldsymbol{V}^\pi$ is the fixed point of $\boldsymbol{B}^\pi$. Therefore, we can write:

$$\boldsymbol{B}^\pi \cdot \boldsymbol{V}^\pi = \boldsymbol{V}^\pi$$