SUPPLEMENT TO PREMIUM CONTROL WITH REINFORCEMENT LEARNING

LINA PALMBORG, FILIP LINDSKOG

1. Policy iteration for the simple model

If the state space is not too large and explicit expressions for the transition probabilities are available, then we may solve both the optimisation problem with a constraint on the action space and with a terminal state numerically using policy iteration. For the simple model, the surplus dynamics is

$$G_{t+1} = G_t + \frac{1}{2}N(P_t + P_{t-1}) - (\beta_0 + \beta_1 N) - \mathrm{PC}_{t+1} + \mathrm{IE}_{t+1},$$

where $\mathrm{IE}_{t+1} = 0$ if $G_t \leq 0$ and otherwise distributed according to (7) in [2]. The constraint (11) in [2] here means that P_t must be sufficiently large to satisfy

$$(1 + \xi \mathbf{1}_{\{G_t > 0\}})G_t + \frac{1}{2}N(P_t + P_{t-1}) - (\beta_0 + \beta_1 N + \mu N) \geq 0.$$

The transition probabilities are given by

$$\begin{aligned}
&\mathrm{P}(S_{t+1} = (k,p) \mid S_t = (g,q), P_t = p) \\
&\quad = \mathrm{P}(\mathrm{IE}_{t+1} + G_t - \mathrm{PC}_{t+1} = k + (\beta_0 + \beta_1 N) - \frac{1}{2}N(p+q) \mid (G_t, P_{t-1}, P_t) = (g,q,p)) \\
&\quad = \begin{cases} \mathrm{P}\left(\mathrm{PC}_{t+1} = g - m\right), & g \leq 0, \\ \sum_{\{l: m+l \geq 0\}} \mathrm{P}(\mathrm{PC}_{t+1} = l)\, \mathrm{P}\left(\mathrm{IE}_{t+1} + G_t = m + l \mid G_t = g\right), & g > 0, \end{cases}
\end{aligned}$$

where $m = k + (\beta_0 + \beta_1 N) - N(p+q)/2$. Due to the truncation of the surplus process, the transition probabilities are adjusted follows: For $g \leq 0$, we have

$$\mathrm{P}(S_{n+1} = (k,p) \mid S_n = (g,q), P_n = p) = \mathrm{P}\left(I_{n+1} = g - k - \beta N + \frac{1}{2}N(p+q)\right).$$

Let $k_{\min}$ and $k_{\max}$ be the minimum and maximum allowed surplus values. Then

$$\begin{aligned}
&\mathrm{P}(S_{n+1} = (k,p) \mid S_n = (g,q), P_n = p) \\
&\quad = \begin{cases} \sum_{l \leq k_{\min}} \mathrm{P}(I_{n+1} = g - l - \beta N + \frac{1}{2}N(p+q)), & k = k_{\min}, \\ \mathrm{P}(I_{n+1} = g - k - \beta N + \frac{1}{2}N(p+q)), & \text{if } k \in (k_{\min}, k_{\max}), \\ \sum_{l \geq k_{\max}} \mathrm{P}(I_{n+1} = g - l - \beta N + \frac{1}{2}N(p+q)), & k = k_{\max}. \end{cases}
\end{aligned}$$

Date: November 30, 2022.

Department of Mathematics, Stockholm University, 106 91 Stockholm, Sweden.

For $g > 0$, we have

$$\begin{aligned}&\mathrm{P}(S_{n+1} = (k,p) \mid S_n = (g,q), P_n = p)\\ &\quad= \sum_{\{l: f(k,p,q)+l \geq 0\}} \mathrm{P}(I_{n+1} = l)\, \mathrm{P}\left(\mathrm{IE}_{n+1} + G_n = f(k,p,q) + l \mid G_n = g\right),\end{aligned}$$

where $f(k,p,q) = k + \beta N - N(p+q)/2$. Let $\widetilde{I}_{n+1}$ be defined as follows:

$$\mathrm{P}(\widetilde{I}_{n+1} = l) = \begin{cases} \mathrm{P}(I_{n+1} = l), & l < l_{\max},\\ \sum_{l \geq l_{\max}} \mathrm{P}(I_{n+1} = l), & l = l_{\max}, \end{cases}$$

where $l_{\max}$ is the the $(1 - 10^{-10})$-quantile of $\mathsf{Pois}(N\mu)$. Then

$$\begin{aligned}&\mathrm{P}(S_{n+1} = (k,p) \mid S_n = (g,q), P_n = p)\\ &= \begin{cases} \sum_{l=0}^{l_{\max}} \mathrm{P}(\widetilde{I}_{n+1} = l) \sum_{k \leq k_{\min}} \mathrm{P}\left(\mathrm{IE}_{n+1} + G_n = f(k,p,q) + l \mid G_n = g\right), & k = k_{\min},\\ \sum_{l=0}^{l_{\max}} \mathrm{P}(\widetilde{I}_{n+1} = l)\, \mathrm{P}\left(\mathrm{IE}_{n+1} + G_n = f(k,p,q) + l \mid G_n = g\right), & k \in (k_{\min}, k_{\max}),\\ \sum_{l=0}^{l_{\max}} \mathrm{P}(\widetilde{I}_{n+1} = l) \sum_{k \geq k_{\max}} \mathrm{P}\left(\mathrm{IE}_{n+1} + G_n = f(k,p,q) + l \mid G_n = g\right), & k = k_{\max}. \end{cases}\end{aligned}$$

2. Q-learning

For Q-learning, the iterative update in search for the optimal action-value function is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha_t\big(R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\big). \tag{1}$$

Given that all state-action pairs continue to be updated, it has been shown in [5] that Q-learning converges to the true optimal action-value function if the step size parameter $0 \leq \alpha_t \leq 1$ satisfies the following stochastic approximation conditions

$$\sum_{k=1}^{\infty} \alpha_{t^k(s,a)} = \infty, \quad \sum_{k=1}^{\infty} \alpha^2_{t^k(s,a)} < \infty, \quad \text{for all } s \in \mathcal{S}, a \in \mathcal{A}(s), \tag{2}$$

where $t^k(s,a)$ is the time step when a visit in state s is followed by taking action a for the kth time.

2.1. Numerical illustration of Q-learning for simple model.

We use the following step size parameter after the kth time a visit in state s is followed by taking action a,

$$\alpha_{t^k(s,a)} = \frac{1}{k^{0.5+\theta}}, \quad \theta = 0.001.$$

This ensures that the stochastic approximation conditions (2) are satisfied, while still allowing for larger step sizes compared to the more standard choice $\alpha_t(s,a) = 1/t$. For the behaviour policy, we set $\varepsilon = 0.2$. The starting state is chosen uniformly at random from the state space. $Q(s,a)$ is initialised to zero for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$ to encourage initial exploration. Since all rewards are negative the true action-value function must be negative for all state-action pairs, hence setting the initial value to zero will encourage that all actions are tried early on. This technique for setting the initial values is called "optimistic initial values" in [4, Ch. 2.6]. To further encourage exploration of the state space, since

discounting will lead to rewards after a large number of steps having a very limited effect on the total reward, we run each episode for at most 100 steps, before resetting to a starting state, again selected uniformly at random from $\mathcal{S}$.

Figure 1 shows the optimal policy for the simplified model using Q-learning. As can be seen in the figures, the Q-learning algorithm has not fully converged to the true optimal policy, despite having been run for a very large number of iterations. This is not too surprising when one considers the fraction of time spent in each state under the optimal policy, see Figure 1 in [2]. The ε-greedy policy and restarting each episode after at most 100 steps ensures that exploration continues when using Q-learning, hence the fraction of time spent in each state during the Q-learning algorithm will not be quite as extreme as in Figure 1 in [2], but there are still many states that will be visited very rarely. Consider e.g. the probability of getting a negative surplus after charging a very high premium (i.e. ending up in the upper right corner of Figure 1 in [2]); the claims payment in the period needs to be quite extreme for this state to be visited, unless the process starts in this state. We have used a step size that guarantees convergence of the algorithm, however, it is possible that a suitably chosen constant step size might lead to the algorithm converging faster.

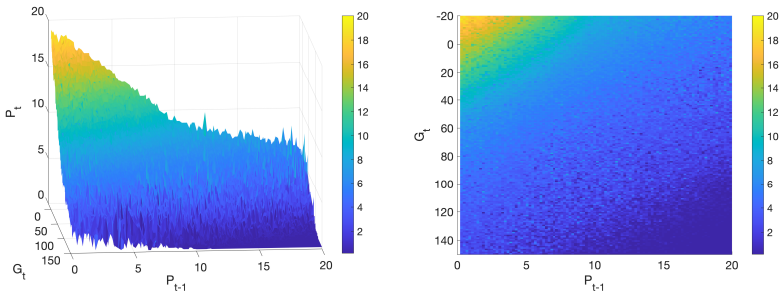

FIGURE 1. Approximate optimal policy for simplified model with terminal state using Q-learning.

3. Softmax policy is Lipschitz continuous

Let $q_{s,a} \in \mathbb{R}$ denote the element of q that correspond to state $s \in \mathcal{S}, a \in \mathcal{A}$, i.e. for the case where we use linear function approximation, $q_{s,a} = w^\top x(s,a)$. Furthermore, let $q_s = (q_{s,a})_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$. An example of a Lipschitz continuous policy is the softmax policy. In this case, the policy improvement operator is given by $\Gamma(q) = (\sigma(q_s))_{s\in\mathcal{S}}$, where σ is the softmax function,

$$\sigma(q_s) = \frac{\exp\left\{\frac{1}{\tau}q_s\right\}}{\sum_{a\in\mathcal{A}}\exp\left\{\frac{1}{\tau}q_{s,a}\right\}}.$$

To see that this policy improvement operator is Lipschitz continuous, first note that the softmax function σ is $1/\tau$-Lipschitz. The softmax function σ is differentiable, hence (see

e.g. [3, Thm 9.19])

$$\|\sigma(q_s) - \sigma(q_s')\|_2 \leq \sup_{q_s} \|D\sigma(q_s)\|_2 \|q_s - q_s'\|_2,$$

where $D\sigma(q_s)$ denotes the Jacobian matrix of σ with respect to q_s, and $\|D\sigma(q_s)\|_2$ is the spectral norm of $D\sigma(q_s)$. Let σ_a denote the component of $\sigma(q_s)$ that corresponds to action $a \in \mathcal{A}$. Then

$$\frac{\partial \sigma_a}{\partial q_{s,a}} = \frac{1}{\tau}\sigma_a(1-\sigma_a), \quad \frac{\partial \sigma_a}{\partial q_{s,a'}} = -\frac{1}{\tau}\sigma_a\sigma_{a'}, \quad \text{for } a \neq a'.$$

It is easy to verify that $D\sigma(q_s)$ is positive semi-definite [1, p. 74]. Hence all eigenvaues of $D\sigma(q_s)$ are non-negative, and

$$\begin{aligned}\|D\sigma(q_s)\|_2 = \lambda_{\max}(D\sigma(q_s)) &\leq \sum_i \lambda_i(D\sigma(q_s)) = \operatorname{tr}(D\sigma(q_s)) = \frac{1}{\tau}\sum_i \sigma_i(1-\sigma_i) \\ &\leq \frac{1}{\tau}\sum_i \sigma_i = \frac{1}{\tau},\end{aligned}$$

where $\lambda_{\max}(D\sigma(q_s))$, $\lambda_i(D\sigma(q_s))$, and $\operatorname{tr}(D\sigma(q_s))$ denote respectively the largest eigenvalue, the ith eigenvalue, and the trace of $D\sigma(q_s)$. Now,

$$\|\Gamma(q) - \Gamma(q')\|_2^2 = \sum_{s\in\mathcal{S}} \|\sigma(q_s) - \sigma(q_s')\|_2^2 \leq \sum_{s\in|\mathcal{S}|} \frac{1}{\tau^2}\|q_s - q_s'\|_2^2 = \frac{1}{\tau^2}\|q - q'\|_2^2,$$

i.e. Γ is Lipschitz continuous with constant $L = 1/\tau$.

4. Benchmark policies

4.1. **Best constant policy.** When using a constant policy p irrespective of state, $S_t = (G_t, p)$ for all $t > 0$, for the simple model. We want to find p that minimises

$$\begin{aligned}\mathrm{E}\left[\sum_{t=0}^{T} \gamma^t h(p, S_{t+1})\right] &= \mathrm{E}\left[\sum_{t=0}^{T-1} \gamma^t c(p) + \gamma^T(1+\eta)c(\max\mathcal{A})\right] \\ &= \mathrm{E}\left[c(p)\frac{1-\gamma^T}{1-\gamma} + \gamma^T(1+\eta)c(\max\mathcal{A})\right] \\ &= \frac{c(p)}{1-\gamma} + \left((1-\eta)c(\max\mathcal{A}) - \frac{c(p)}{1-\gamma}\right)\mathrm{E}[\gamma^T],\end{aligned} \tag{3}$$

where $\mathrm{E}[\gamma^T] = \sum_{t=1}^{\infty} \mathrm{P}(T=t)\gamma^t$ and

$$\mathrm{P}(T \leq t) = \sum_{s'\in\mathcal{S}^+\setminus\mathcal{S}} \mathrm{P}(S_t = s') = \mathrm{P}(G_t < \min\mathcal{G}).$$

Since the state space is finite, we can label the states $0, 1, \ldots, |\mathcal{G}||\mathcal{A}|$ (where state 0 represents all terminal (absorbing) states. Let $P = (p_{ij} : i, j \in \{0, 1, \ldots, |\mathcal{G}||\mathcal{A}|\})$, where

$p_{ij} = \mathrm{P}(S_t = j \mid S_{t-1} = i)$, and $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_{|\mathcal{G}||\mathcal{A}|})^\top$, where $\lambda_j = \mathrm{P}(S_0 = j)$ and $\lambda_0 = 0$. Then

$$\mathrm{P}(G_t = k) = \sum_j (\lambda^\top P^t)_{\{j:G_t=k\}}.$$

Based on this we can compute (3) for each p (computing $\mathrm{E}[\gamma^T]$ by truncating the sum at some large value) from which we determine that $p = 7.4$ minimises (3), for the simple model.

For the intermediate and realistic model, we are not able to compute the best constant policy as above due to the dimension of the state space. Instead we can simulate the total expected discounted reward per episode for different values of the constant policy p. Based on these simulations we can again conclude that $p = 7.4$ minimises (3) also for the intermediate model. For the realistic model $p = 11.5$ minimises (3).

4.2. **Myopic policy for MDP with constraint.** The myopic policy is the policy that maximises immediate (next-step) rewards. For the model with a constraint on the action space, the myopic policy is the solution to the following optimisation problem

$$\underset{p}{\text{minimise}}\ \mathrm{E}[c(p) \mid S_0 = s, P_0 = p] \quad \text{subject to}\ \mathrm{E}[G_1 \mid S_0 = s, P_0 = p] \geq 0.$$

Since c is an increasing function, it is easy to compute the myopic policy; it is given by the lowest premium level that satisfies the constraint. For the simple model we have

$$G_{t+1} = G_t + \frac{1}{2}N(P_t + P_{t-1}) - \beta_1 N - \beta_0 - \mathrm{PC}_{t+1} + \mathrm{IE}_{t+1}.$$

Hence, for each $s = (g, q)$, we need to find the lowest premium level $p = \pi(g, q)$ that satisfies

$$(1 + \xi \mathbf{1}_{\{g>0\}})g + \frac{1}{2}N(p + q) - (\beta_0 + (\beta_1 + \mu)N) \geq 0,$$

hence

$$\pi(g, q) = \min\Big\{p \in \mathcal{A} : p \geq 2\Big(\beta_1 + \mu + \frac{\beta_0 - (1 + \xi \mathbf{1}_{\{g>0\}})g}{N} - q\Big)\Big\}.$$

The myopic policy for the MDP with a constraint on the action space can be seen in Figure 2 for the simple model.

For the intermediate model we have

$$\begin{aligned} G_{t+1} = G_t &+ \frac{1}{2}(N_{t+1}P_t + N_t P_{t-1}) - (\alpha_2\mu + \beta_1)\frac{1}{2}(N_{t+1} + N_t) + \alpha_2\mu\frac{1}{2}(N_t + N_{t-1}) - \beta_0 \\ &- \mathrm{PC}_{t+1} + \mathrm{IE}_{t+1}. \end{aligned}$$

Hence, for each $s = (g, q, n_0, n_{-1})$, we want to find the lowest premium level $p = \pi(g, q, n_0, n_{-1})$ that satisfies

$$(1 + \xi \mathbf{1}_{\{g>0\}})g + \frac{1}{2}ap^b(p - \beta_1 - \mu) + \frac{1}{2}n_0(q - \beta_1 - \mu) - \beta_0 \geq 0. \tag{4}$$

Note that $\pi(g, q, n_0, n_{-1})$ does not depend on n_{-1}. Let $\mathcal{P}(g, q, n_0)$ be the set of premium levels such that (4) is satisfied. Note that for our choice of $\mathcal{A}$ and $\mathcal{S}$, there exist

$(g, q, n_0, n_{-1}) \in \mathcal{S}$ such that $\mathcal{P}(g, q, n_0) \cap \mathcal{A} = \emptyset$. Hence we let the myopic policy for the intermediate model with a constraint on the action space be given by

$$\pi(g, q, n_0, n_{-1}) = \begin{cases} \min\{p \in \mathcal{A} : p \in \mathcal{P}(g, q, n_0)\}, & \mathcal{P}(g, q, n_0) \cap \mathcal{A} \neq \emptyset, \\ \max \mathcal{A}, & \mathcal{P}(g, q, n_0) \cap \mathcal{A} = \emptyset, \end{cases}$$

i.e. if no $p \in \mathcal{A}$ satisfies the constraint, then the maximum premium level is chosen.

For the realistic model we have

$$\begin{aligned} G_{t+1} = G_t + \frac{1}{2}(N_{t+1}P_t + N_tP_{t-1}) - \beta_1\frac{1}{2}(N_{t+1} + N_t) - \beta_0 \\ + \mathrm{IE}_{t+1} - \mathrm{IC}_{t+1} + \mathrm{RP}_{t+1}. \end{aligned}$$

Hence, for each $s = (g, q, n_0, n_{-1}, c_1, c_2, \ldots, c_9)$, we want to find the lowest premium level $p = \pi(g, q, n_0, n_{-1}, c_1, c_2, \ldots, c_9)$ that satisfies

$$(1 + \xi \mathbf{1}_{\{g>0\}})g + \frac{1}{2}ap^b\left(p - \beta_1 - c_0\prod_{k=1}^{J-1} f_k\right) + \frac{1}{2}n_0\left(q - \beta_1 - c_0\prod_{k=1}^{J-1} f_k\right) - \beta_0 \geq 0. \tag{5}$$

Note that $\pi(g, q, n_0, n_{-1}, c_1, c_2, \ldots, c_9)$ does not depend on $n_{-1}, c_1, c_2, \ldots, c_9$.

4.3. **Myopic policy for MDP with terminal state.** For the model with a terminal state, for each state $s \in \mathcal{S}$ we want to find $p \in \mathcal{A}$ that minimises

$$\begin{aligned} \mathrm{E}[h(p, S_1) \mid S_0 = s, P_0 = p] = c(p)\,\mathrm{P}(G_1 \geq \min \mathcal{G} \mid S_0 = s, P_0 = p) \\ + c(\max \mathcal{A})(1 + \eta)\,\mathrm{P}_\pi(G_1 < \min \mathcal{G} \mid S_0 = s, P_0 = p), \end{aligned} \tag{6}$$

hence for this model the myopic policy is not quite as easy to compute as for the case when we have a constraint on the action space, since we now need to determine

$$\mathrm{P}(G_1 \geq \min \mathcal{G} \mid S_0 = s, P_0 = p) = \sum_{k=\min \mathcal{G}}^{\infty} \mathrm{P}(G_1 = k \mid S_0 = s, P_0 = p)$$

instead of the expectation in (20) in [2]. From Section 1 we see that for the simple model

$$\begin{aligned} &\mathrm{P}(G_1 = k \mid S_0 = (g, q), P_0 = p) \\ &= \begin{cases} \mathrm{P}\left(\mathrm{PC}_1 = g - m\right), & g \leq 0, \\ \sum_{\{l: m+l \geq 0\}} \mathrm{P}(\mathrm{PC}_1 = l)\,\mathrm{P}\left(\mathrm{IE}_1 + G_0 = m + l \mid G_0 = g\right), & g > 0, \end{cases} \end{aligned}$$

where $m = k + (\beta_0 + \beta_1 N) - N(p + q)/2$. Based on this we can compute the expectation in (6) for each premium level. The myopic policy for the MDP with a terminal state can be seen in Figure 2 for the simple model.

5. Details for the realistic model

To analyse the difference between the approximate optimal policy and the best benchmark policy, we simulate 300 episodes for a few different starting states, two of which can be seen in Figure 3. For both starting states, $C_{-1,1} = c_0 \cdot 2 \cdot 10^5$, and $C_{-j,j} = c_0 \cdot 2 \cdot 10^5 \prod_{k=1}^{j-1} f_k$ for $j = 2, \ldots, 9$. Note that each star in the figures correspond to one or more terminations at that time point. The total number of terminations (of 300 episodes) are:

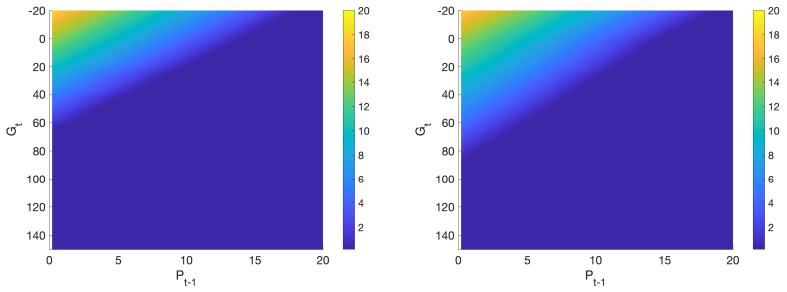

FIGURE 2. Left: myopic policy for simple model with constraint. Right: myopic policy for simplified model with terminal state.

j	0	1	2	3	4	5	6	7	8	9
$\widehat{\mu}_j$	13.249	0.662	0.199	0.111	0.075	0.039	0.027	0.014	0.016	0.001
$\widehat{\nu}_j$	0.154	0.096	0.067	0.016	0.022	0.017	0.012	0.004	0.008	0
$\widehat{f}_j$	-	1.947	1.223	1.118	1.078	1.040	1.028	1.015	1.016	1.001
$\widehat{\alpha}_{j+1}$	0.0316	0.300	0.137	0.089	0.066	0.036	0.026	0.014	0.016	0.001

TABLE 1. Parameter estimates for the model for the cumulative claims payments in (8) in [2].

j	1	2	3	4	5	6	7	8	9
$\min(C_{t,j})$	0.234	0.416	0.489	0.546	0.586	0.608	0.624	0.633	0.642
$\max(C_{t,j})$	2.094	4.417	5.596	6.265	6.777	7.063	7.266	7.372	7.493

TABLE 2. Truncation of cumulative claims payments, in 10^6.

$S_0 = (-3 \cdot 10^5, 3, 2 \cdot 10^5, 2 \cdot 10^5, C_{-1,1}, \ldots, C_{-9,9})$: Fourier 3: 6, interval policy: 67, best constant: 257. $S_0 = (3 \cdot 10^6, 21, 1.75 \cdot 10^5, 1.75 \cdot 10^5, C_{-1,1}, \ldots, C_{-9,9})$: Fourier 3: 1, interval policy: 0, best constant: 0. Comparing the two policies, we see that the interval policy on average tends to charge a lower premium, and the fluctuation in the premium level over time appears to be comparable. However, for the strained starting state $S_0 = (-3 \cdot 10^5, 3, 2 \cdot 10^5, 2 \cdot 10^5, C_{-1,1}, \ldots, C_{-9,9})$, the interval policy leads to many more immediate terminations (67 instead of 6 for the approximate policy), and in the starting state with a much higher surplus, $S_0 = (3 \cdot 10^6, 21, 1.75 \cdot 10^5, 1.75 \cdot 10^5, C_{-1,1}, \ldots, C_{-9,9})$, the approximate policy outperforms the interval policy, despite having one termination at approximately year 30, due to the higher initial premium level set by the interval policy, and the higher variation in the premium level at the first few time steps for the interval policy.

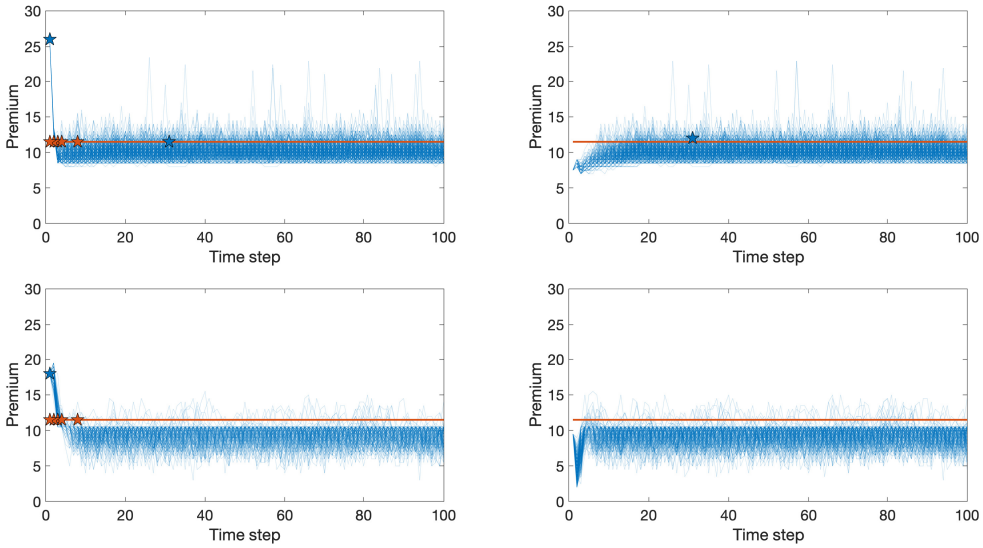

FIGURE 3. Realistic model. First row: policy with 3rd order Fourier basis. Second row: interval policy. Left: starting state $S_0 = (-3 \cdot 10^5, 3, 2 \cdot 10^5, 2 \cdot 10^5, C_{-1,1}, \ldots, C_{-9,9})$. Right: starting state $S_0 = (3 \cdot 10^6, 21, 1.75 \cdot 10^5, 1.75 \cdot 10^5, C_{-1,1}, \ldots, C_{-9,9})$. The red line shows the best constant policy. Each star indicates at least one termination at that time step.

REFERENCES

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
[2] Lina Palmborg and Filip Lindskog. Premium control with reinforcement learning. 2022.
[3] Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.
[4] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
[5] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.