Homework 4 - Solutions

1. Let $w^* := \arg\min_{w\in\mathcal{K}} \sum_{t=1}^T f_t(w)$. Since f_t is a convex function, we have

$$f_t(w_t) - f_t(w^*) = \langle \nabla f_t(w_t), w_t - w^* \rangle = \langle \mathbb{E}[g_t \mid w_t], w_t - w^* \rangle .$$

Let $y_t := w_{t-1} - \eta g_{t-1}$ (hence $w_t = \Pi_{\mathcal{K}}(y_t)$), and thus

$$\begin{aligned} f_t(w_t) - f_t(w^*) &= \frac{1}{2\eta}\mathbb{E}[2(w_t - y_{t+1})^T(w_t - w^*) \mid w_t] \\ &= \frac{1}{2\eta}\left(\|w_t - w^*\|^2 + \mathbb{E}[\|w_t - y_{t+1}\|^2 - \|w^* - y_{t+1}\|^2 \mid w_t]\right) \\ &= \frac{1}{2\eta}\left(\|w_t - w^*\|^2 + \mathbb{E}[\eta^2\|g_t\|^2 - \|w^* - y_{t+1}\|^2 \mid w_t]\right). \end{aligned}$$

But $\|w^* - y_{t+1}\|^2 \geq \|w^* - w_{t+1}\|^2$ since $w_{t+1} = \Pi_{\mathcal{K}}(y_{t+1})$, and $\mathcal{K}$ is convex. Thus

$$f_t(w_t) - f_t(w^*) \leq \frac{1}{2\eta}\left(\|w_t - w^*\|^2 + \mathbb{E}[\eta^2\|g_t\|^2 - \|w^* - w_{t+1}\|^2 \mid w_t]\right).$$

Taking expectation w.r.t. w_t on both sides, we have

$$\mathbb{E}[f_t(w_t) - f_t(w^*)] \leq \frac{1}{2\eta}\mathbb{E}\left[\|w_t - w^*\|^2 + \eta^2\|g_t\|^2 - \|w^* - w_{t+1}\|^2\right].$$

Summing over $t = 1, 2, \ldots, T$, we have

$$\mathbb{E}\left[\sum_{t=1}^T (f_t(w_t) - f_t(w^*))\right] \leq \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^T \|g_t\|^2\right] + \frac{\mathbb{E}[\|w_1 - w^*\|^2]}{2\eta} \leq \frac{\eta}{2}TG^2 + \frac{D^2}{2\eta}$$

where the last inequality follows since $\|g_t\| \leq G$ w.p. 1, $w_1 = 0$, and $D = \max_{w\in\mathcal{K}} \|x\|$.

2. Let us choose states 0 and 1, and choose actions "LEFT" and "RIGHT" for both the states. The state transition diagram is as in Figure 1. The thick lines in the figure corresponds to the action "LEFT", and the dotted lines, to the action "RIGHT". Thus the state transition matrices for the actions look like

$$T_{LEFT} = \begin{pmatrix} .9 & .1 \\ .8 & .2 \end{pmatrix}, \quad T_{RIGHT} = \begin{pmatrix} .3 & .7 \\ .1 & .9 \end{pmatrix}.$$

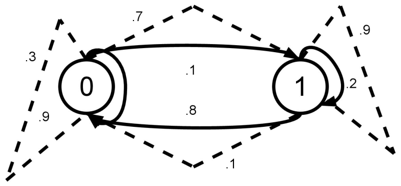

Figure 1: State Transition Diagram

Now, let $R(0, \text{LEFT}) = 1$, $R(0, \text{RIGHT}) = 3$, $R(1, \text{LEFT}) = 2$, $R(1, \text{RIGHT}) = 4$. The value iteration equation for this MDP is

$$\begin{aligned}
V_2(0) &= \max_{a \in \{\text{LEFT, RIGHT}\}} R(0,a) = R(0, \text{RIGHT}) = 3\\
V_2(1) &= \max_{a \in \{\text{LEFT, RIGHT}\}} R(1,a) = R(1, \text{RIGHT}) = 4\\
V_1(0) &= \max_{a \in \{\text{LEFT, RIGHT}\}} \big(R(0,a) + \sum_{s \in \{0,1\}} T(0,a,s)V_2(s)\big) = \max(4.1, 6.7) = 6.7\\
V_1(1) &= \max_{a \in \{\text{LEFT, RIGHT}\}} \big(R(1,a) + \sum_{s \in \{0,1\}} T(1,a,s)V_2(s)\big) = \max(5.2, 7.9) = 7.9\\
V_0(0) &= \max_{a \in \{\text{LEFT, RIGHT}\}} \big(R(0,a) + \sum_{s \in \{0,1\}} T(0,a,s)V_1(s)\big) = \max(7.82, 10.54) = 10.54\\
V_0(1) &= \max_{a \in \{\text{LEFT, RIGHT}\}} \big(R(1,a) + \sum_{s \in \{0,1\}} T(0,a,s)V_1(s)\big) = \max(8.94, 11.78) = 11.78.
\end{aligned}$$

Thus the optimal value function $V_0^*(0) = 10.54$, and $V_0^*(1) = 11.78$. The optimal policy is to choose “RIGHT” always.

3. (a) When $\beta \to 0$, we have $\exp(\beta \sum_{i=1}^N r_i) \approx 1 + \sum_{i=1}^N r_i$. Thus maximizing $J_\beta^\pi(s)$ is equivalent to maximizing $\text{sgn}(\beta)\mathbb{E}_\pi[1 + \sum_{i=1}^N r_i]$, which in turn is equivalent to maximizing the reward-to-go-function $V_k^\pi(s)$ defined in class. Thus the optimal policy for low values of β is just

$$\pi_k(s) \in \arg\max_a \left(R(s,a) + \sum_{s' \in \mathcal{S}} T(s,a,s')V_{k+1}^*(s') \right).$$

(b) Define the Reward-to-go function as $V_{\beta,k}^\pi(s) := \mathbb{E}_\pi[\text{sgn}(\beta)\exp(\beta \sum_{i=k+1}^N r_i) \mid \mathcal{S}_k = s]$. Let the optimal value function $V_{\beta,k}^*(s) = \max_\pi V_{\beta,k}^\pi(s)$. On the same lines of the proof shown in class, we will show that

$$V_{\beta,k}^*(s) = \max_a \left(\text{sgn}(\beta)\exp(\beta R(s,a)) \left(\sum_{s' \in \mathcal{S}} T(s,a,s')V_{\beta,k+1}^*(s') \right) \right). \tag{1}$$

We will show (1) by mathematical induction. The equation is true when $k = N$, since $V_{\beta,N}^*(s) = 1$ for any $s \in \mathcal{S}$. Now let it be true for every $k \geq (l+1)$. We will show first that $V_{\beta,k}^*(s)$ is at most the RHS of (1), i.e., $V_{\beta,k}^\pi(s)$ is at most the RHS of (1) for any tail policy $\pi^{(k)} \equiv (\pi_k, \ldots, \pi_{N-1})$. For any deterministic policy π, we have

$$\begin{aligned}
V_{\beta,k}^\pi(s) &= \mathbb{E}_\pi \left[\text{sgn}(\beta)\exp\left(\beta \sum_{i=k+1}^N r_i \right) \mid S_k = s \right]\\
&= \text{sgn}(\beta)\exp(\beta R(s,\pi_k(s))) \left(\sum_{s' \in \mathcal{S}} T(s,\pi_k(s),s')\mathbb{E}_\pi \left[\sum_{i=k+2}^N r_i \mid S_{k+1} = s' \right] \right)\\
&= \text{sgn}(\beta)\exp(\beta R(s,\pi_k(s))) \left(\sum_{s' \in \mathcal{S}} T(s,\pi_k(s),s')V_{\beta,k+1}^\pi(s') \right).
\end{aligned}$$

Thus for a tail policy $\pi^{(k)}$, we have

$$\begin{aligned}
V_{\beta,k}^{\pi^{(k)}}(s) &= \text{sgn}(\beta)\exp(\beta R(s,\pi_k^{(k)}(s))) \left(\sum_{s' \in \mathcal{S}} T(s,\pi_k^{(k)}(s),s')V_{\beta,k+1}^{\pi^{(k)}}(s') \right)\\
&\leq \max_a \left(\text{sgn}(\beta)\exp(\beta R(s,a)) \left(\sum_{s' \in \mathcal{S}} T(s,a,s')V_{\beta,k+1}^{\pi^{(k)}}(s') \right) \right)\\
&\leq \max_a \left(\text{sgn}(\beta)\exp(\beta R(s,a)) \left(\sum_{s' \in \mathcal{S}} T(s,a,s')V_{\beta,k+1}^*(s') \right) \right).
\end{aligned}$$

To show $V^*_{\beta,k}(s)$ is at least the RHS of (1), we only need to exhibit a policy $\pi^{(k)}$ s.t. $V^{\pi^{(k)}}_{\beta,k}$ equals the RHS of (1). At time k, choose action

$$a' \in \arg\max_a [\operatorname{sgn}(\beta)\exp(\beta R(s,a))(\sum_{s'\in\mathcal{S}} T(s,a,s')V^*_{\beta,k+1}(s'))].$$

Using this policy clearly makes $V^{\pi^{(k)}}_{\beta,k}$ equal to the RHS of (1). Now, for $\beta > 0$, we have,

$$v^*_{\beta,k}(s) := \log V^*_{\beta,k}(s) = \max_a \left(\beta R(s,a) + \log\left(\sum_{s'\in\mathcal{S}} T(s,a,s')\exp(v^*_{\beta,k+1}(s'))\right)\right),$$

and this differs from the standard case by the existence of a "log" term.

(c) Let the random variable X be defined as $X = \sum_{i=1}^N r_i$. Let $\mathbb{P}(X = r_j) = p_j > 0, j = 1,\dots,M$, with $r_1 < r_2 < \dots < r_M$. Then the term $J^\pi_\beta(s)$, when $\beta \to \infty$, can be written as

$$J^\pi_\beta(s) = \operatorname{sgn}(\beta)\mathbb{E}_\pi[\exp(\beta\sum_{i=1}^N r_i)] = \sum_{j=1}^M P_j \exp(\beta r_j) \approx P_M \exp(\beta r_M).$$

Thus we need to choose a policy that maximizes the maximum sum of rewards, with maximum overall probability. In other words, the chosen policy π must maximize x for which $\mathbb{P}_\pi[\max X = x] > 0$. In case of a tie between two policies π' and π'' giving the same value of x, the policy that maximizes $\mathbb{P}_{\pi\in\{\pi',\pi''\}}[\max X = x]$ should be chosen. Let $V^\pi_{\infty,k}(s)$ be such a maximum (i.e., equal to x) when $S_k = s$, and let $V^*_{\infty,k}(s) := \max_\pi V^\pi_{\infty,k}(s)$. It is easy to derive the following recursive equation for $V^*_{\infty,k}(s)$:

$$V^*_{\infty,k}(s) = \max_a \left(R(s,a) + \max_{s':\mathbf{1}(T(s,a,s')>0)} V^*_{\infty,k+1}(s')\right).$$

We define the set $S^*_k := \arg\max_{s':\mathbf{1}(T(s,a,s')>0)} V^*_{\infty,k+1}(s')$, and the quantity $T^*_{\infty,k}(s) := \max_{s'\in S^*_k} T(s,a,s')$. Then, the optimal policy is a deterministic policy π that satisfies $\pi_k(s) \in \arg\max_a(R(s,a) + \max_{s':\mathbf{1}(T(s,a,s')>0)} V^*_{\infty,k+1}(s'))$, and breaks ties in favour of the policy having a higher value of $\Pi^N_{j=k+1} T^*_{\infty,j}(s)$.

Similarly the term $J^\pi_\beta(s)$, when $\beta \to -\infty$, can be written as

$$J^\pi_\beta(s) = \operatorname{sgn}(\beta)\mathbb{E}_\pi[\exp(\beta\sum_{i=1}^N r_i)] = -\sum_{j=1}^M P_j \exp(\beta r_j) \approx -P_1 \exp(\beta r_1).$$

Thus we need to choose a policy that maximizes the minimum sum of rewards, with minimum overall probability. In other words, the chosen policy π must maximize x for which $\mathbb{P}_\pi[\min X = x] > 0$. In case of a tie between two policies π' and π'' giving the same value of x, the policy that minimizes $\mathbb{P}_{\pi\in\{\pi',\pi''\}}[\min X = x]$ should be chosen. Let $V^\pi_{-\infty,k}(s)$ be such a minimum (i.e., equal to x) when $S_k = s$, and let $V^*_{-\infty,k}(s) := \max_\pi V^\pi_{-\infty,k}(s)$. It is easy to derive the following recursive equation for $V^*_{-\infty,k}(s)$:

$$V^*_{-\infty,k}(s) = \max_a \left(R(s,a) + \min_{s':\mathbf{1}(T(s,a,s')>0)} V^*_{-\infty,k+1}(s')\right).$$

We define the set $S^*_k := \arg\min_{s':\mathbf{1}(T(s,a,s')>0)} V^*_{-\infty,k+1}(s')$, and the quantity $T^*_{-\infty,k}(s) := \min_{s'\in S^*_k} T(s,a,s')$. Then, the optimal policy is a deterministic policy π that satisfies $\pi_k(s) \in \arg\max_a(R(s,a) + \min_{s':\mathbf{1}(T(s,a,s')>0)} V^*_{-\infty,k+1}(s'))$, and breaks ties in favour of the policy having a lower value of $\Pi^N_{j=k+1} T^*_{-\infty,j}(s)$.

4. Let us consider the sequence $(.9, .99, .999, .9999, .99999)$ for the sequence of γ approaching 1. Number of value iterations were roughly $(135, 1350, 13500, 135000, 1350000)$, with a total running time of $(0.05, 0.5, 5, 50, 500)$ seconds respectively. The policy iteration had roughly 3 iterations, and had a total running time to be less than 10 milliseconds, for any value of γ. Clearly, the number of value iterations move proportional to $\frac{1}{1-\gamma}$, but from simulations, it seems that the number of policy iterations don't depend on γ.