Online Supplement for Exploiting the Structural Properties of the Underlying Markov Decision Problem in the Q-Learning Algorithm

Sumit Kunnumkal, Huseyin Topaloglu
School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853, USA,
{summit@orie.cornell.edu, topaloglu@orie.cornell.edu}

In this document, we prove Lemmas 2 and 5. Our proof of Lemma 2 uses the following generalization of the supermartingale convergence theorem.

Proposition 7 *Let $\{X_t\}, \{Y_t\}$ and $\{Z_t\}$ be sequences of positive random variables adopted to the filtration $\{\mathcal{G}_t\}$. Assume that $\mathbb{E}\{X_{t+1} \,|\, \mathcal{G}_t\} \leq X_t - Y_t + Z_t$ for all $t = 0, 1, \ldots$ and $\sum_{t=0}^{\infty} Z_t < \infty$ w.p.1. Then, $\{X_t\}$ converges to a positive random variable and we have $\sum_{t=0}^{\infty} Y_t < \infty$ w.p.1.*

We now prove Lemma 2.

Proof of Lemma 2 Letting R_t and $G_t \in \mathbb{R}^{n \times n}$ be the diagonal matrices whose diagonal components are respectively $\{\rho_t(i) : i \in \mathcal{S}\}$ and $\{\gamma_t(i) : i \in \mathcal{S}\}$, (9) can be written in vector notation as $\hat{y}_t = y_t + R_t\, G_t\big[\, y^* + \theta_t - y_t\big]$. By the nonexpansiveness of the projection operator $\Pi^{L,U}$, we have

$$\|y_{t+1} - y^*\|_2^2 \leq \|\hat{y}_t - y^*\|_2^2 = \|y_t - y^*\|_2^2 - 2\big\langle y_t - y^*, R_t\, G_t\big[\, y_t - y^*\big]\big\rangle + 2\big\langle y_t - y^*, R_t\, G_t\, \theta_t\big\rangle + \|R_t\, G_t\big[\, y^* + \theta_t - y_t\big]\|_2^2. \quad (34)$$

Letting P_t be the diagonal matrix whose diagonal components are $\{\mathbb{P}\{i_t = i \,|\, \mathcal{G}_t\} : i \in \mathcal{S}\}$, and noting the fact that y_t is $\mathcal{G}_t$-measurable and $\mathbb{E}\big\{\rho_t(i)\ \gamma_t(i) \,|\, \mathcal{G}_t\big\} = \mathbb{P}\{i_t = i \,|\, \mathcal{G}_t\}\ \gamma_t(i)$, we have

$$\mathbb{E}\big\{\big\langle y_t - y^*, R_t\, G_t\big[\, y_t - y^*\big]\big\rangle \,|\, \mathcal{G}_t\big\} = \big\langle y_t - y^*, \mathbb{E}\big\{R_t\, G_t \,|\, \mathcal{G}_t\big\}[y_t - y^*]\big\rangle = \big\langle y_t - y^*, P_t\, G_t\big[\, y_t - y^*\big]\big\rangle. \quad (35)$$

Noting (10) and (14), we have

$$\mathbb{E}\big\{\big\langle y_t - y^*, R_t\, G_t\, \theta_t\big\rangle \,|\, \mathcal{G}_t, i_t\big\} = \mathbb{E}\big\{\big[\, y_t(i_t) - y^*(i_t)\big]\gamma_t(i_t)\, \theta_t(i_t) \,|\, \mathcal{G}_t, i_t\big\} = \big[\, y_t(i_t) - y^*(i_t)\big]\gamma_t(i_t)\, \mathbb{E}\big\{\theta_t(i_t) \,|\, \mathcal{G}_t, i_t\big\} = 0 \qquad \text{w.p.1},$$

which implies that

$$\mathbb{E}\big\{\big\langle y_t - y^*, R_t\, G_t\, \theta_t\big\rangle \,|\, \mathcal{G}_t\big\} = \mathbb{E}\Big\{\mathbb{E}\big\{\big\langle y_t - y^*, R_t\, G_t\, \theta_t\big\rangle \,|\, \mathcal{G}_t, i_t\big\} \,|\, \mathcal{G}_t\Big\} = 0 \qquad \text{w.p.1.} \tag{36}$$

Since y_t is the projection of $\hat{y}_{t-1}$ onto $\mathcal{P}^{L,U}$, we have $y_t \in \mathcal{P}^{L,U}$. Then, (14), (15) and the fact that $y^* \in \mathcal{P}^{L,U}$ imply that

$$\begin{aligned}
\mathbb{E}\big\{\|R_t\, G_t\big[\, y^* + \theta_t - y_t\big]\|_2^2 \,|\, \mathcal{G}_t, i_t\big\} &= \mathbb{E}\big\{\|R_t\, G_t\big[\, y^* - y_t\big]\|_2^2 \,|\, \mathcal{G}_t, i_t\big\} \\
&\quad + 2\,\mathbb{E}\big\{\big\langle R_t\, G_t\big[\, y^* - y_t\big], R_t\, G_t\, \theta_t\big\rangle \,|\, \mathcal{G}_t, i_t\big\} + \mathbb{E}\big\{\|R_t\, G_t\, \theta_t\|_2^2 \,|\, \mathcal{G}_t, i_t\big\} \\
&= [\gamma_t(i_t)]^2\big[\, y^*(i_t) - y_t(i_t)\big]^2 + 2\,[\gamma_t(i_t)]^2\big[\, y^*(i_t) - y_t(i_t)\big]\,\mathbb{E}\big\{\theta_t(i_t) \,|\, \mathcal{G}_t, i_t\big\} + [\gamma_t(i_t)]^2\,\mathbb{E}\big\{[\theta_t(i_t)]^2 \,|\, \mathcal{G}_t, i_t\big\} \\
&\leq [\gamma_t(i_t)]^2\,\big\{4\,[|L| \vee |U|]^2 + A\big\} \qquad \text{w.p.1,}
\end{aligned}$$

where we use $a \vee b = \max\{a, b\}$. Taking the expectation of the expression above conditional on $\mathcal{G}_t$, we have

$$\mathbb{E}\big\{\|R_t\, G_t\big[\, y^* + \theta_t - y_t\big]\|_2^2 \,|\, \mathcal{G}_t\big\} \leq \mathbb{E}\big\{[\gamma_t(i_t)]^2 \,|\, \mathcal{G}_t\big\}\,\big\{4\,[|L| \vee |U|]^2 + A\big\} \qquad \text{w.p.1.} \tag{37}$$

Taking the expectation of (34) conditional on $\mathcal{G}_t$ and using (35)-(37), we obtain

$$\begin{aligned}
&\mathbb{E}\big\{\|y_{t+1} - y^*\|_2^2 \,|\, \mathcal{G}_t\big\} \\
&\quad \leq \|y_t - y^*\|_2^2 - 2\big\langle y_t - y^*, P_t\, G_t\big[\, y_t - y^*\big]\big\rangle + \mathbb{E}\big\{[\gamma_t(i_t)]^2 \,|\, \mathcal{G}_t\big\}\,\big\{4\,[|L| \vee |U|]^2 + A\big\}\,.
\end{aligned} \tag{38}$$

On the other hand, using (13) and appealing to the monotone convergence theorem, we have

$$\begin{aligned}
\mathbb{E}\Big\{\sum_{t=0}^{\infty}\mathbb{E}\big\{[\gamma_t(i_t)]^2 \,|\, \mathcal{G}_t\big\}\Big\} &= \sum_{t=0}^{\infty}\mathbb{E}\big\{[\gamma_t(i_t)]^2\big\} \\
&= \sum_{t=0}^{\infty}\mathbb{E}\Big\{\sum_{i=1}^{n}\rho_t(i)\,[\gamma_t(i)]^2\Big\} = \sum_{i=1}^{n}\sum_{t=0}^{\infty}\mathbb{E}\big\{\rho_t(i)\,[\gamma_t(i)]^2\big\} < \infty,
\end{aligned}$$

which implies that $\sum_{t=0}^{\infty}\mathbb{E}\big\{[\gamma_t(i_t)]^2 \,|\, \mathcal{G}_t\big\} < \infty$ w.p.1. Therefore, noting (38), we can use Proposition 7 to conclude that the sequence $\{\|y_t - y^*\|_2^2\}$ converges w.p.1 and

$$\sum_{t=0}^{\infty}\big\langle y_t - y^*, P_t\, G_t\big[\, y_t - y^*\big]\big\rangle < \infty \qquad \text{w.p.1.}$$

Then, we have $\sum_{t=0}^{\infty}[\min_{i\in\mathcal{S}}\gamma_t(i)]\,\big\langle y_t - y^*, P_t\big[\, y_t - y^*\big]\big\rangle \leq \sum_{t=0}^{\infty}\big\langle y_t - y^*, P_t\, G_t\big[\, y_t - y^*\big]\big\rangle < \infty$ w.p.1. Therefore, (11) and (12) imply that there exists a subset of iterations $\mathcal{T}$ such that the sequence $\{y_t - y^*\}_{t\in\mathcal{T}}$ converges to 0 w.p.1. Since $\{\|y_t - y^*\|_2^2\}$ converges w.p.1, the whole sequence $\{y_t - y^*\}$ converges to 0 w.p.1. $\square$

We now turn to Lemma 5. Letting $v(0) = L$ and $v(n+1) = U$, the result of the projection $\Pi^{L,U}\hat{y}$ is the optimal solution to the problem

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i=1}^{n}[v(i) - \hat{y}(i)]^2 \\ \text{subject to} \quad & v(i) \leq v(i+1) \qquad \text{for all } i \in \{0, \ldots, n\}. \end{aligned} \tag{39}$$

Associating the nonnegative Lagrange multipliers $\{\lambda(i) : i \in \{0, \ldots, n\}\}$ with the constraints, the Karush-Kuhn-Tucker conditions for this problem are

$$\hat{y}(i) - \lambda(i) + \lambda(i-1) = v(i) \qquad \text{for all } i \in \{1, \ldots, n\} \tag{40}$$

$$\lambda(i)\,[v(i+1) - v(i)] = 0 \qquad \text{for all } i \in \{0, \ldots, n\}. \tag{41}$$

The following result is useful when proving Lemma 5.

Lemma 8 *Assume that $y \in \mathcal{P}^{L,U}$. Fix $i^* \in \mathcal{S}$ and let $\hat{y} \in \mathbb{R}^n$ be obtained by*

$$\hat{y}(i) = \begin{cases} y(i) + \alpha\,[A - y(i)] & \text{if } i = i^* \\ y(i) & \text{otherwise,} \end{cases}$$

where A is a scalar and $\alpha \in [0, 1]$. Let $v = \Pi^{L,U}\hat{y}$ be computed by solving problem (39) *and $\{\lambda(i) : i \in \{0, \ldots, n\}\}$ be the corresponding optimal Lagrange multipliers. Then, one of the following cases holds.*

1) *Either we have $\lambda(i) = 0$ for all $i \in \{0, \ldots, n\}$.*

2) *Or there exist κ and $\mu \in \{0, \ldots, n\}$ with $\kappa \leq \mu$ such that $\lambda(i) > 0$ if and only if $i \in \{\kappa, \ldots, \mu\}$.*

Proof of Lemma 8 To reach a contradiction, we assume that there exist $\kappa_1, \mu_1, \kappa_2, \mu_2, \ldots, \kappa_k$ and $\mu_k \in \{0, \ldots, n\}$ with $k \geq 2$ such that $\kappa_1 < \mu_1 + 1 < \kappa_2 < \mu_2 + 1 < \ldots < \kappa_k < \mu_k + 1$ and $\lambda(i) > 0$ if and only if $i \in \{\kappa_1, \ldots, \mu_1\} \cup \{\kappa_2, \ldots, \mu_2\} \cup \ldots \cup \{\kappa_k, \ldots, \mu_k\}$. We consider four cases.

Case 1 – Assume that $\kappa_1 > 0$ and $\mu_2 < n$. Since $\lambda(i) > 0$ for all $i \in \{\kappa_1, \ldots, \mu_1\}$, (41) implies that $v(\kappa_1) = v(\kappa_1 + 1) = \ldots = v(\mu_1 + 1)$. Since $\lambda(\kappa_1) > 0$, $\lambda(\mu_1) > 0$ and $\lambda(\kappa_1 - 1) = \lambda(\mu_1 + 1) = 0$, (40) implies that $v(\kappa_1) = \hat{y}(\kappa_1) - \lambda(\kappa_1) < \hat{y}(\kappa_1)$ and $v(\mu_1 + 1) = \hat{y}(\mu_1 + 1) + \lambda(\mu_1) > \hat{y}(\mu_1 + 1)$. Therefore, we obtain, $\hat{y}(\kappa_1) > v(\kappa_1) = v(\mu_1 + 1) > \hat{y}(\mu_1 + 1)$. On the other hand, since $y \in \mathcal{P}^{L,U}$, we have $y(\kappa_1) \leq y(\mu_1 + 1)$. Therefore, since $\hat{y}$ differs from y only in the i^*-th component, we must have $i^* = \kappa_1$ or $i^* = \mu_1 + 1$. By using a similar

argument, we can obtain $\hat{y}(\kappa_2) > \hat{y}(\mu_2 + 1)$, which implies that we must also have $i^* = \kappa_2$ or $i^* = \mu_2 + 1$. Since we cannot have both $i^* \in \{\kappa_1, \mu_1 + 1\}$ and $i^* \in \{\kappa_2, \mu_2 + 1\}$, we reach a contradiction.

Case 2 – Assume that $\kappa_1 = 0$ and $\mu_2 < n$. Since $\lambda(i) > 0$ for all $i \in \{0, \ldots, \mu_1\}$ and $v(0) = L$, (41) implies that $L = v(1) = v(2) = \ldots = v(\mu_1 + 1)$. Then, since $\lambda(\mu_1) > 0$ and $\lambda(\mu_1 + 1) = 0$, (40) implies that $L = v(\mu_1 + 1) = \hat{y}(\mu_1 + 1) + \lambda(\mu_1) > \hat{y}(\mu_1 + 1)$. On the other hand, since $y \in \mathcal{P}^{L,U}$, we have $y(\mu_1 + 1) \geq L$. Therefore, since $\hat{y}$ differs from y only in the i^*-th component, we must have $i^* = \mu_1 + 1$. Following an argument similar to the one in the previous case, we must also have $i^* = \kappa_2$ or $i^* = \mu_2 + 1$. Since we cannot have both $i^* = \mu_1 + 1$ and $i^* \in \{\kappa_2, \mu_2 + 1\}$, we reach a contradiction.

The cases where we have $\kappa_1 > 0$ and $\mu_2 = n$, or $\kappa_1 = 0$ and $\mu_2 = n$ can be handled using similar arguments. □

We now prove Lemma 5.

Proof of Lemma 5 Let v be computed by solving problem (39) and $\{\lambda(i) : i \in \{0, \ldots, n\}\}$ be the corresponding optimal Lagrange multipliers. We show that there exists $\ell \in \mathcal{S}$ that satisfies (23)-(25). Before we begin, we note that $v \in \mathcal{P}^{L,U}$ since v is a feasible solution to problem (39).

If $\lambda(i) = 0$ for all $i \in \{0, \ldots, n\}$, then (40) implies that $v(i) = \hat{y}(i)$ for all $i \in \{1, \ldots, n\}$. In this case, we set $\ell = i^*$. Since $v \in \mathcal{P}^{L,U}$, we have $L \leq v(i^*) = \hat{y}(i^*) \leq U$, which implies that $v(i^*) = \hat{y}(i^*) = [\hat{y}(i^*) \wedge U] \vee L$. Therefore, (23) is satisfied. Since $\ell = i^*$, (24) and (25) do not need to be shown.

We now assume that there exists at least one strictly positive Lagrange multiplier. Let κ and μ with $\kappa \leq \mu$ be such that $\lambda(i) > 0$ if and only if $i \in \{\kappa, \ldots, \mu\}$. The existence of such κ and μ is guaranteed by Lemma 8. We consider four cases.

Case 1 – Assume that $\kappa > 0$ and $\mu < n$. In this case, following an argument similar to the one in Case 1 in the proof of Lemma 8, we obtain $v(\kappa) = v(\kappa + 1) = \ldots = v(\mu + 1)$ and we must have $i^* = \kappa$ or $i^* = \mu + 1$. By adding (40) for all $i \in \{\kappa, \ldots, \mu + 1\}$, we obtain

$$v(\kappa) = v(\kappa + 1) = \ldots = v(\mu + 1) = \frac{\hat{y}(\kappa) + \ldots + \hat{y}(\mu + 1)}{\mu - \kappa + 2}. \tag{42}$$

Since $\lambda(i) = 0$ for all $i \notin \{\kappa, \ldots, \mu\}$, (40) implies that

$$v(i) = \hat{y}(i) \qquad \text{for all } i \notin \{\kappa, \ldots, \mu + 1\}. \tag{43}$$

Finally, since $\lambda(\kappa) > 0$, $\lambda(\mu) > 0$ and $\lambda(\kappa - 1) = \lambda(\mu + 1) = 0$, (40) implies that $v(\kappa) = \hat{y}(\kappa) - \lambda(\kappa) < \hat{y}(\kappa)$ and $v(\mu + 1) = \hat{y}(\mu + 1) + \lambda(\mu) > \hat{y}(\mu + 1)$. Then, using (42), we obtain

$$\hat{y}(\mu + 1) < v(\mu + 1) = \frac{\hat{y}(\kappa) + \ldots + \hat{y}(\mu + 1)}{\mu - \kappa + 2} = v(\kappa) < \hat{y}(\kappa). \tag{44}$$

Since we must have $i^* = \kappa$ or $i^* = \mu + 1$, we consider the following two subcases.

Case 1.a – Assume that $i^* = \kappa$. In this case, we set $\ell = \mu + 1$ and we have $\ell > i^*$. Since $v \in \mathcal{P}^{L,U}$, (42) implies that

$$v(i^*) = v(i^* + 1) = \ldots = v(\ell) = \frac{\hat{y}(i^*) + \ldots + \hat{y}(\ell)}{\ell - i^* + 1} \leq U,$$

from which we obtain

$$v(i) = \frac{\hat{y}(i^*) + \ldots + \hat{y}(\ell)}{\ell - i^* + 1} \wedge U \qquad \text{for all } i \in \{i^*, \ldots, \ell\}.$$

By (43), we also have $v(i) = \hat{y}(i)$ for all $i \notin \{i^*, \ldots, \ell\}$. Therefore, (23) is satisfied. Since $y \in \mathcal{P}^{L,U}$ and $\hat{y}$ differs from y only in the i^*-th component, we have $\hat{y}(i^* + 1) \leq \hat{y}(i^* + 2) \leq \ldots \leq \hat{y}(\ell) = \hat{y}(\mu + 1)$. Then, noting (44) shows that (24) is satisfied. Since $\ell > i^*$, (25) does not need to be shown.

Case 1.b – Assume that $i^* = \mu + 1$. In this case, we set $\ell = \kappa$ and we have $\ell < i^*$. Since $v \in \mathcal{P}^{L,U}$, (42) implies that

$$v(\ell) = v(\ell + 1) = \ldots = v(i^*) = \frac{\hat{y}(\ell) + \ldots + \hat{y}(i^*)}{i^* - \ell + 1} \geq L,$$

from which we obtain

$$v(i) = \frac{\hat{y}(\ell) + \ldots + \hat{y}(i^*)}{i^* - \ell + 1} \vee L \qquad \text{for all } i \in \{\ell, \ldots, i^*\}.$$

By (43), we also have $v(i) = \hat{y}(i)$ for all $i \notin \{\ell, \ldots, i^*\}$. Therefore, (23) is satisfied. Since $y \in \mathcal{P}^{L,U}$ and $\hat{y}$ differs from y only in the i^*-th component, we have $\hat{y}(\kappa) = \hat{y}(\ell) \leq \hat{y}(\ell + 1) \leq \ldots \leq \hat{y}(i^* - 1)$. Then, noting (44) shows that (25) is satisfied. Since $\ell < i^*$, (24) does not need to be shown.

Case 2 – Assume that $\kappa = 0$ and $\mu < n$. In this case, following an argument similar to the one in Case 2 in the proof of Lemma 8, we obtain $L = v(1) = v(2) = \ldots = v(\mu + 1)$, $\hat{y}(\mu + 1) < L$ and we must have $i^* = \mu + 1$. By adding (40) for all $i \in \{1, \ldots, \mu + 1\}$, we obtain

$$L = v(1) = v(2) = \ldots = v(\mu + 1) = \frac{\hat{y}(1) + \ldots + \hat{y}(\mu + 1) + \lambda(0)}{\mu + 1} > \frac{\hat{y}(1) + \ldots + \hat{y}(\mu + 1)}{\mu + 1}. \tag{45}$$

Since $\lambda(i) = 0$ for all $i \notin \{0, \ldots, \mu\}$, (40) implies that

$$v(i) = \hat{y}(i) \qquad \text{for all } i \notin \{1, \ldots, \mu + 1\}. \tag{46}$$

Since we must have $i^* = \mu + 1$, we consider the following two subcases.

Case 2.a – Assume that $\mu > 0$. In this case, we have $i^* > 1$, we set $\ell = 1$ and we obtain $\ell < i^*$. By (45), we have

$$v(i) = L = \frac{\hat{y}(\ell) + \ldots + \hat{y}(i^*)}{i^* - \ell + 1} \vee L \qquad \text{for all } i \in \{\ell, \ldots, i^*\}.$$

By (46), we also have $v(i) = \hat{y}(i)$ for all $i \notin \{\ell, \ldots, i^*\}$. Therefore, (23) is satisfied. Since $y \in \mathcal{P}^{L,U}$ and $\hat{y}$ differs from y only in the i^*-th component, we have $L \leq \hat{y}(\ell) \leq \hat{y}(\ell + 1) \leq \ldots \leq \hat{y}(i^* - 1)$. We also have $\hat{y}(i^*) = \hat{y}(\mu + 1) < L$. Therefore, we have $\hat{y}(i^*) < \hat{y}(\ell) \leq \hat{y}(\ell + 1) \leq \ldots \leq \hat{y}(i^* - 1)$, which implies that $\hat{y}(i^*) < [\hat{y}(\ell) + \ldots + \hat{y}(i^*)]/[i^* - \ell + 1]$. Noting (45), we obtain

$$\hat{y}(i^*) < \frac{\hat{y}(\ell) + \ldots + \hat{y}(i^*)}{i^* - \ell + 1} < L \leq \hat{y}(\ell) \leq \ldots \leq \hat{y}(i^* - 1),$$

which shows that (25) is satisfied. Since $\ell < i^*$, (24) does not need to be shown.

Case 2.b – Assume that $\mu = 0$. In this case, we have $i^* = 1$, we set $\ell = 1$ and we obtain $\ell = i^*$. Since $\hat{y}(1) = \hat{y}(\mu + 1) < L \leq U$, (45) implies that $v(i^*) = v(1) = L = [\hat{y}(1) \wedge U] \vee L = [\hat{y}(i^*) \wedge U] \vee L$. By (46), we also have $v(i) = \hat{y}(i)$ for all $i \notin \{i^*\}$. Therefore, (23) is satisfied. Since $\ell = i^*$, (24) and (25) do not need to be shown.

The cases where we have $\kappa > 0$ and $\mu = n$, or $\kappa = 0$ and $\mu = n$ can be handled using similar arguments. □