A. Average Losses

Lemma 4 in Appendix C motivates the following optimization problem:

$$\min_h v^\top (Lh - h)\,, \tag{17}$$

where v is a distribution over the state space. If $\widehat{h}$ is an ϵ-optimal solution, then

$$v^\top (L\widehat{h} - \widehat{h}) \leq v^\top (Lh - h) + \epsilon\ .$$

Thus, by Lemma 4 in Appendix C,

$$\lambda_{\pi_{\widehat{h}}} + (v - \mu_{\pi_{\widehat{h}}})^\top (L\widehat{h} - \widehat{h}) \leq \lambda_{\pi_h} + (v - \mu_{\pi_h})^\top (Lh - h) + \epsilon\ .$$

Thus, for any $\widehat{\lambda}$ and λ,

$$\lambda_{\pi_{\widehat{h}}} + (v - \mu_{\pi_{\widehat{h}}})^\top (L\widehat{h} - \widehat{h} - \widehat{\lambda}\mathbf{1}) \leq \lambda_{\pi_h} + (v - \mu_{\pi_h})^\top (Lh - h - \lambda\mathbf{1}) + \epsilon\ .$$

Thus,

$$\lambda_{\pi_{\widehat{h}}} - \lambda_{\pi_h} \leq \left\|v - \mu_{\pi_{\widehat{h}}}\right\|_1 \left\|L\widehat{h} - \widehat{h} - \widehat{\lambda}\mathbf{1}\right\|_\infty + \left\|Lh - h - \lambda\mathbf{1}\right\|_{1,v} + \left\|Lh - h - \lambda\mathbf{1}\right\|_{1,\mu_{\pi_h}}\ .$$

Unfortunately, the optimization objective (17) is not convex.

B. Proofs of Section 2

Before proving Theorem 1, we prove a useful lemma.

Lemma 3. *Let $J : \mathcal{X} \to \mathbb{R}$ be a function. We have that*

$$J_{P_J}(x_1) - J(x_1) = \sum_{T \in \mathcal{T}} P_J(T) \sum_{x \in T} (LJ - J)(x)\ .$$

Proof. We have that

$$\begin{aligned} \ell(x, P_J) &= q(x) + \sum_{x' \in \mathcal{X}} P_J(x, x') \log \frac{P_J(x, x')}{P_0(x, x')} \\ &= q(x) - \sum_{x' \in \mathcal{X}} P_J(x, x') J(x') - \log Z(x)\ . \end{aligned} \tag{18}$$

By definition and (18),

$$J_{P_J}(x) = q(x) + \sum_{x' \in \mathcal{X}} P_J(x, x')(J_{P_J}(x') - J(x')) - \log Z(x)\ .$$

Thus,

$$\begin{aligned} J_{P_J}(x) - J(x) &= q(x) + \sum_{x' \in \mathcal{X}} P_J(x, x')(J_{P_J}(x') - J(x')) - \log Z(x) - J(x) \\ &= (LJ - J)(x) + \sum_{x' \in \mathcal{X}} P_J(x, x')(J_{P_J}(x') - J(x'))\ . \end{aligned}$$

Let $f(x) = J_{P_J}(x) - J(x)$ and $g(x) = (LJ - J)(x)$ so that $f(x) = g(x) + \sum_{x' \in \mathcal{X}} P_J(x, x') f(x')$. Because there are no loops and there exists an absorbing state such that $(LJ - J)(z) = 0$, we obtain the desired result:

$$J_{P_J}(x_1) - J(x_1) = \sum_{T \in \mathcal{T}} P_J(T) \sum_{x \in T} (LJ - J)(x)\ .$$

□

Proof of Theorem 1. Because $\widehat{w} \in \mathcal{W}$, by the positivity assumption below (7) we have that $J_{\widehat{w}}(x) \le -\log g$ for any state x. Thus,

$$
\begin{aligned}
(LJ_{\widehat{w}})(x) &= q(x) - \log \sum_{x'} P_0(x,x') e^{-J_{\widehat{w}}(x')} \\
&\le q(x) + \sum_{x'} P_0(x,x') \left(-\log e^{-J_{\widehat{w}}(x')} \right) \\
&\le Q - \log g \; .
\end{aligned}
$$

Thus, for any x,

$$
\max\{J_{\widehat{w}}(x), (LJ_{\widehat{w}})(x)\} \le Q - \log g \; . \tag{19}
$$

By the fact that $\widehat{w}$ is an ϵ-optimal solution, for any $w \in \mathcal{W}$, we have

$$
\begin{aligned}
&J_{\widehat{w}}(x_1) + H \sum_{T\in\mathcal{T}} v(T) \sum_{x\in T} \left| \Psi(x,:)\widehat{w} - e^{-q(x)} P_0(x,:)\Psi\widehat{w} \right| \le \\
&\qquad J_w(x_1) + H \sum_{T\in\mathcal{T}} v(T) \sum_{x\in T} \left| \Psi(x,:)w - e^{-q(x)} P_0(x,:)\Psi w \right| + \epsilon \; .
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&J_{\widehat{w}}(x_1) + H e^{-Q+\log g} \sum_{T\in\mathcal{T}} v(T) \sum_{x\in T} \left| J_{\widehat{w}}(x) - LJ_{\widehat{w}}(x) \right| \le \\
&\qquad J_w(x_1) + H \sum_{T\in\mathcal{T}} v(T) \sum_{x\in T} e^{-l_w}(x) \left| J_w(x) - LJ_w(x) \right| + \epsilon \, ,
\end{aligned}
$$

where we used (9) and (19). Thus, by the choice of H and Lemma 3,

$$
\begin{aligned}
J_{P_{J_{\widehat{w}}}}(x_1) + \sum_{T\in\mathcal{T}} (v(T) - P_{J_{\widehat{w}}}(T)) \sum_{x\in T} \left| J_{\widehat{w}}(x) - LJ_{\widehat{w}}(x) \right| &\le J_{P_{J_w}}(x_1) \\
&\quad + H \sum_{T\in\mathcal{T}} v(T) \sum_{x\in T} e^{-l_w}(x) \left| J_w(x) - LJ_w(x) \right| \\
&\quad + \sum_{T\in\mathcal{T}} P_{J_w}(T) \sum_{x\in T} \left| J_w(x) - LJ_w(x) \right| + \epsilon \; .
\end{aligned}
$$

Thus,

$$
\begin{aligned}
J_{P_{J_{\widehat{w}}}}(x_1) - J_{P_{J_w}}(x_1) &\le \sum_{T\in\mathcal{T}} (P_{J_{\widehat{w}}}(T) - v(T)) \sum_{x\in T} \left| J_{\widehat{w}}(x) - LJ_{\widehat{w}}(x) \right| \\
&\quad + H \sum_{T\in\mathcal{T}} v(T) \sum_{x\in T} e^{-l_w}(x) \left| J_w(x) - LJ_w(x) \right| \\
&\quad + \sum_{T\in\mathcal{T}} P_{J_w}(T) \sum_{x\in T} \left| J_w(x) - LJ_w(x) \right| + \epsilon \\
&\le \left\| P_{J_{\widehat{w}}} - v \right\|_1 \max_{T\in\mathcal{T}} \sum_{x\in T} \left| J_{\widehat{w}}(x) - LJ_{\widehat{w}}(x) \right| \\
&\quad + H \sum_{T\in\mathcal{T}} v(T) \sum_{x\in T} e^{-l_w}(x) \left| J_w(x) - LJ_w(x) \right| \\
&\quad + \sum_{T\in\mathcal{T}} P_{J_w}(T) \sum_{x\in T} \left| J_w(x) - LJ_w(x) \right| + \epsilon \; .
\end{aligned}
$$

□

Input: Starting state x_1, number of rounds N, a decreasing sequence of step sizes (η_t), a positive v over states, estimate of optimal average cost b.
Let $\Pi_{\mathcal{W}}$ be the Euclidean projection onto $\mathcal{W}$.
Initialize $w_1 = \mathbf{0}$.
for $t := 1, 2, \dots, N$ **do**
  Sample state $x \sim v/\|v\|_1$.
  Compute subgradient estimate r_t defined by (20).
  Update $w_{t+1} = \Pi_{\mathcal{W}}(w_t - \eta_t r_t)$.
end for
$\widehat{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$.
Return policy $P_{h_{\widehat{w}_T}}$ defined in Section 3.

Figure 3. The Stochastic Subgradient Method for Average Cost Markov Decision Processes

C. Algorithm and Proofs of Section 3

The stochastic subgradient algorithm for average cost MDPs is presented in Figure 3, where the stochastic subgradient of $c(w)$ for a randomly sampled state x takes the following form,

$$r(w) = \|v\|_1 \text{sign}\left(e^{-b}\Psi(x,:)w - e^{-q(x)}P_0(x,:)\Psi w\right)\left(e^{-b}\Psi(x,:) - e^{-q(x)}P_0(x,:)\Psi\right) . \tag{20}$$

Before proving Theorem 2, we prove a useful lemma.

Lemma 4. *Let $h : \mathcal{X} \to \mathbb{R}$ be a bounded function and assume that the Markov chain induced by the greedy policy P_h is irreducible and aperiodic. Then, we have that*

$$\lambda_{P_h} = \mu_{P_h}^{\top}(Lh - h) ,$$

where μ_{P_h} is the stationary distribution with respect to P_h.

Proof. The proof argument uses ideas from the proof of Theorem 8.4.1 in Puterman (1994). Let $f(x) = (Lh)(x) - h(x)$. We have that

$$P_h f = P_h \ell(:, P_h) + P_h^2 h - P_h h = P_h \ell(:, P_h) + P_h(P_h - I)h .$$

By repeating this argument, we get that $P_h^s f = P_h^s \ell(., P_h) + P_h^s(P_h - I)h$. Summing over $s = 1 \dots t$, we obtain

$$\sum_{s=1}^{t} P_h^s f = \sum_{s=1}^{t} P_h^s \ell(:, P_h) + (P_h^t - I)h .$$

Averaging and taking the limit, we obtain

$$P_h^{\infty} f = \lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} P_h^s f = \lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} P_h^s \ell(:, P_h) + \lim_{t\to\infty} \frac{1}{t}(P_h^t - I)h = \lambda_{P_h}\mathbf{1} ,$$

where we used $\lambda_P \mathbf{1} = P^{\infty}\ell(:, P)$ and boundedness of $P_h^{\infty} h$. Thus, $\lambda_{P_h} = \mu_{P_h}^{\top} f$. □

Proof of Theorem 2. For a differential value function h, let $V = e^{-h}$. We know that $Lh = q - \log Z$ and $Z(x) = \sum_{x'} P_0(x, x')e^{-h(x')} = \sum_{x'} P_0(x, x')V(x')$. Then

$$e^{-Lh} - e^{-h-b} = e^{-q+\log Z} - e^{-h-b} = e^{-q}P_0 V - e^{-b}V .$$

Let $\widehat{w}$ be an ϵ-optimal solution, then for any $w \in \mathcal{W}$, we have,

$$v^{\top}\left|e^{-q}P_0\Psi\widehat{w} - e^{-b}\Psi\widehat{w}\right| \le v^{\top}\left|e^{-q}P_0\Psi w - e^{-b}\Psi w\right| + \epsilon .$$

Recall that $h_{\widehat{w}} = -\log \Psi \widehat{w}$. Let $u_{\widehat{w}} = \max(Lh_{\widehat{w}}, h_{\widehat{w}} + b)$ and $l_w = \min(Lh_w, h_w + b)$. By (9),

$$\left(e^{-u_{\widehat{w}}} \odot v\right)^\top |Lh_{\widehat{w}} - h_{\widehat{w}} - b| \leq \left(e^{-l_w} \odot v\right)^\top |Lh_w - h_w - b| + \epsilon \,.$$

By Lemma 4, we have

$$\lambda_{P_{h_{\widehat{w}}}} - b \leq \mu^T_{P_{h_{\widehat{w}}}} |Lh_{\widehat{w}} - h_{\widehat{w}} - b| \,,$$

which further implies that,

$$-b + \lambda_{P_{h_{\widehat{w}}}} + (e^{-u_{\widehat{w}}} \odot v - \mu_{P_{h_{\widehat{w}}}})^\top |Lh_{\widehat{w}} - h_{\widehat{w}} - b| \leq \left(e^{-l_w} \odot v\right)^\top |Lh_w - h_w - b| + \epsilon \,.$$

This gives the performance bound in the theorem,

$$\begin{aligned}\lambda_{P_{h_{\widehat{w}}}} - \lambda_{P_{h_w}} \leq \left|b - \lambda_{P_{h_w}}\right| + \|(e^{-u_{\widehat{w}}} \odot v - \mu_{P_{h_{\widehat{w}}}})\|_1 \|Lh_{\widehat{w}} - h_{\widehat{w}} - b\|_\infty \\ + \|\left(Lh_w - h_w - b\right)\|_{1,(e^{-l_w} \odot v)} + \epsilon.\end{aligned}$$

□