1 Proof of Theorem 1

Proof Let us rewrite the minimization problem,

$$W^{t+1} \quad = \quad \arg\min_{W}(\langle \bar{G}^t, W\rangle + \lambda\sqrt{K}\sum_{i}^{d}||W_{\cdot i}||_2 + \frac{\beta_t}{2t}\sum_{i}^{d}||W_{\cdot i}||_2^2)$$

Since the minimization problem is component-wise on one column of W, we can focus on each of the column of W separately to find its solution.

$$W_{\cdot i}^{t+1} = \arg\min_{W_{\cdot i}}\left(\langle \bar{G}_{\cdot i}^t, W_{\cdot i}\rangle + \lambda\sqrt{K}||W_{\cdot i}||_2 + \frac{\beta}{2t}||W_{\cdot i}||_2^2\right)$$

Since inner product of 2 vectors of same length will have smallest value when the 2 vectors are in opposite direction, solution to the above minimization problem should be $W_{\cdot i}^{t+1} = \varphi\bar{G}_{\cdot i}^t$ where $\varphi \leq 0$. We now need to solve the following minimization problem,

$$\varphi \quad = \quad \arg\min_{\varphi\leq 0}\left(\varphi||\bar{G}_{\cdot i}^t||_2^2 - \varphi\lambda\sqrt{K}||\bar{G}_{\cdot i}^t||_2 + \varphi^2\frac{\beta}{2t}||\bar{G}_{\cdot i}^t||_2^2\right)$$

Solving for the minimum point of that familiar quadratic function, we have

$$\varphi = \begin{cases} \frac{t}{\beta_t}\left(\frac{\lambda\sqrt{K}}{||\bar{G}_{\cdot i}^t||_2} - 1\right) & \text{if } ||\bar{G}_{\cdot i}^t||_2 > \lambda\sqrt{K}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the update rule is as in theorem 1.

2 Proof of Theorem 3

Proof For any action a, consider the following expression, where x is a vector of all state attributes and features extracted from state s and e is the action effect leading to s' from s.

$$\begin{aligned}
&\sum_{s'\in S} P^{M_1}(s'|s)\log\left(\frac{P^{M_1}(s'|s)}{P^{M_2}(s'|s)}\right) \\
=&\sum_{e\in E} P^{M_1}(e|s)\log\left(\frac{P^{M_1}(e|s)}{P^{M_2}(e|s)}\right) \\
\leq&\max_{e\in E}\log\left(\frac{P^{M_1}(e|s)}{P^{M_2}(e|s)}\right) \\
=&\max_{e\in E}\log\left(\frac{\exp(W_e^{M_1}x)/\sum_{e'\in E}\exp(W_{e'}^{M_1}x)}{\exp(W_e^{M_2}x)/\sum_{e'\in E}\exp(W_{e'}^{M_2}x)}\right) \\
=&\max_{e\in E}\left[(W_e^{M_1} - W_e^{M_2})x - \log\left(\frac{\sum_{e'\in E}\exp(W_{e'}^{M_1}x)}{\sum_{e'\in E}\exp(W_{e'}^{M_2}x)}\right)\right] \\
\leq&\max_{e\in E}\left[(W_e^{M_1} - W_e^{M_2})x - \log\left(\min_{e'\in E}\left(\frac{\exp(W_{e'}^{M_1}x)}{\exp(W_{e'}^{M_2}x)}\right)\right)\right] \\
=&\max_{e\in E}\left[(W_e^{M_1} - W_e^{M_2})x - \min_{e'\in E}\left(\log\left(\frac{\exp(W_{e'}^{M_1}x)}{\exp(W_{e'}^{M_2}x)}\right)\right)\right] \\
=&\max_{e\in E}\left[(W_e^{M_1} - W_e^{M_2})x - \min_{e'\in E}\left((W_{e'}^{M_1} - W_{e'}^{M_2})x\right)\right] \\
\leq&\max_{e\in E}[||W_e^{M_1} - W_e^{M_2}||_1\sup_s||x(s)||_1 + \max_{e'\in E}(||W_{e'}^{M_1} - W_{e'}^{M_2}||_1\sup_s||x(s)||_1)] \\
\leq&2\max_{e\in E}(||W_e^{M_1} - W_e^{M_2}||_1\sup_s||x(s)||_1)
\end{aligned}$$

The first step is from definition of *effect*. The second step is from the fact that weighted average of elements must be smaller than the largest one. The sixth step is from the property that if a_i and b_i are non-negative, then $\left(\sum_i a_i\right) / \left(\sum_i b_i\right) \geq \min_i(a_i/b_i)$. The seventh step is from monotonicity of logarithmic function.

By Pinsker's inequality,

$$\begin{aligned}&\sum_{s'\in S} P^{M_1}(s'|s)\log\left(\frac{P^{M_1}(s'|s)}{P^{M_2}(s'|s)}\right)\\ \geq&\frac{1}{2}(\sum_{s'\in S}|P^{M_1}(s'|s)-P^{M_2}(s'|s)|)^2\end{aligned}$$

which implies

$$\begin{aligned}&\sum_{s'\in S}|P^{M_1}(s'|s)-P^{M_2}(s'|s)|\\ \leq&2\sqrt{\max_{e\in E}(||W_e^{M_1}-W_e^{M_2}||_1\sup_s||x(s)||_1)}\end{aligned}$$

Extending to all actions,

$$\begin{aligned}&\sum_{s'\in S}|P^{M_1}(s'|s)-P^{M_2}(s'|s)|\\ \leq&\max_{a\in A}\left(2\sqrt{\max_{e\in E}(||W_e^{(a),M_1}-W_e^{(a),M_2}||_1\sup_s||x(s)||_1)}\right)\\ =&2\sqrt{\max_{a\in A,e\in E}(||W_e^{(a),M_1}-W_e^{(a),M_2}||_1\sup_s||x(s)||_1)}\end{aligned}$$

To complete the theorem, the following lemma (see lemma 33 in [1]) is used without proof.

Lemma 1 *Let $M_1=(S,A,P^{M_1},R), M_2=(S,A,P^{M_2},R)$ be two MDPs, and fixed discount factor γ. π_1 and π_2 are their optimal policies respectively. Let V_π^M be the value functions of π in MDP M. If*

$$\sum_{s'\in S}|P^{M_1}-P^{M_2}|(s'|s,a)\leq\epsilon$$

for every state-action (s,a), then $|V_{\pi_2}^{M_1}(s)-V_{\pi_2}^{M_2}(s)|\leq\frac{\gamma V_{max}\epsilon}{1-\gamma}$ and $|V_{\pi_1}^{M_2}(s)-V_{\pi_1}^{M_1}(s)|\leq\frac{\gamma V_{max}\epsilon}{1-\gamma}$, for every $s\in S$.

It is clear that

$$\begin{aligned}&\max_{s\in S}\left(V_{\pi_2}^{M_2}-V_{\pi_1}^{M_2}\right)\\ =&\max_{s\in S}\left(V_{\pi_2}^{M_2}-V_{\pi_1}^{M_1}+V_{\pi_1}^{M_1}-V_{\pi_1}^{M_2}\right)\\ \leq&\max_{s\in S}\left(V_{\pi_2}^{M_2}-V_{\pi_2}^{M_1}+V_{\pi_1}^{M_1}-V_{\pi_1}^{M_2}\right)\\ \leq&\max_{s\in S}|V_{\pi_2}^{M_2}-V_{\pi_2}^{M_1}|+\max_{s\in S}|V_{\pi_1}^{M_1}-V_{\pi_1}^{M_2}|\\ \leq&\frac{2\gamma V_{max}\epsilon}{1-\gamma}.\end{aligned}$$

The proof is therefore complete.

References

[1] Li, L.: A unifying framework for computational reinforcement learning theory. Ph.D. thesis, Rutgers, The State of University of New Jersey (2009)