A Relevant proofs

A.1 Proof of Theorem 1

Proof. The proof is based on Borkar's Theorem for general stochastic approximation recursions with two time scales (Borkar 1997).

A new one-step linear TD solution is defined as:

$$0 = \mathbb{E}[(\delta - \mathbb{E}[\delta])\phi] = -A\theta + b.$$

Thus, the VMTD's solution is $\theta_{\text{VMTD}} = A^{-1}b$.

First, note that recursion (5) can be rewritten as

$$\theta_{k+1} \leftarrow \theta_k + \beta_k \xi(k),$$

where

$$\xi(k) = \frac{\alpha_k}{\beta_k}(\delta_k - \omega_k)\phi_k$$

Due to the settings of step-size schedule $\alpha_k = o(\beta_k)$, $\xi(k) \to 0$ almost surely as $k \to \infty$. That is the increments in iteration (4) are uniformly larger than those in (5), thus (4) is the faster recursion. Along the faster time scale, iterations of (4) and (5) are associated to ODEs system as follows:

$$\dot{\theta}(t) = 0, \tag{A-1}$$

$$\dot{\omega}(t) = \mathbb{E}[\delta_t|\theta(t)] - \omega(t). \tag{A-2}$$

Based on the ODE (A-20), $\theta(t) \equiv \theta$ when viewed from the faster timescale. By the Hirsch lemma (Hirsch 1989), it follows that $||\theta_k - \theta|| \to 0$ a.s. as $k \to \infty$ for some θ that depends on the initial condition θ_0 of recursion (5). Thus, the ODE pair (A-20)-(A-21) can be written as

$$\dot{\omega}(t) = \mathbb{E}[\delta_t|\theta] - \omega(t). \tag{A-3}$$

Consider the function $h(\omega) = \mathbb{E}[\delta|\theta] - \omega$, i.e., the driving vector field of the ODE (A-22). It is easy to find that the function h is Lipschitz with coefficient -1. Let $h_\infty(\cdot)$ be the function defined by $h_\infty(\omega) = \lim_{x \to \infty} \frac{h(x\omega)}{x}$. Then $h_\infty(\omega) = -\omega$, is well-defined. For (A-22), $\omega^* = \mathbb{E}[\delta|\theta]$ is the unique globally asymptotically stable equilibrium. For the ODE

$$\dot{\omega}(t) = h_\infty(\omega(t)) = -\omega(t), \tag{A-4}$$

apply $\vec{V}(\omega) = (-\omega)^\top(-\omega)/2$ as its associated strict Liapunov function. Then, the origin of (A-23) is a globally asymptotically stable equilibrium.

Consider now the recursion (**??**). Let $M_{k+1} = (\delta_k - \omega_k) - \mathbb{E}[(\delta_k - \omega_k)|\mathcal{F}(k)]$, where $\mathcal{F}(k) = \sigma(\omega_l, \theta_l, l \le k; \phi_s, \phi'_s, r_s, s < k)$, $k \ge 1$ are the sigma fields generated by $\omega_0, \theta_0, \omega_{l+1}, \theta_{l+1}, \phi_l, \phi'_l$, $0 \le l < k$. It is easy to verify that $M_{k+1}, k \ge 0$ are integrable random variables that satisfy $\mathbb{E}[M_{k+1}|\mathcal{F}(k)] = 0$, $\forall k \ge 0$. Because ϕ_k, r_k, and ϕ'_k have uniformly bounded second moments, it can be seen that for some constant $c_1 > 0$, $\forall k \ge 0$,

$$\mathbb{E}[||M_{k+1}||^2|\mathcal{F}(k)] \le c_1(1 + ||\omega_k||^2 + ||\theta_k||^2).$$

Now Assumptions (A1) and (A2) of (Borkar and Meyn 2000) are verified. Furthermore, Assumptions (TS) of (Borkar and Meyn 2000) is satisfied by our conditions on the step-size sequences α_k, β_k. Thus, by Theorem 2.2 of (Borkar and Meyn 2000) we obtain that $||\omega_k - \omega^*|| \to 0$ almost surely as $k \to \infty$.

Consider now the slower time scale recursion (5). Based on the above analysis, (5) can be rewritten as

$$\theta_{k+1} \leftarrow \theta_k + \alpha_k(\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k.$$

Let $\mathcal{G}(k) = \sigma(\theta_l, l \le k; \phi_s, \phi'_s, r_s, s < k)$, $k \ge 1$ be the sigma fields generated by $\theta_0, \theta_{l+1}, \phi_l, \phi'_l$, $0 \le l < k$. Let $Z_{k+1} = Y_t - \mathbb{E}[Y_t|\mathcal{G}(k)]$, where

$$Y_k = (\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k.$$

Consequently,

$$\begin{array}{rcl} \mathbb{E}[Y_t|\mathcal{G}(k)] & = & \mathbb{E}[(\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k|\mathcal{G}(k)] \\ & = & \mathbb{E}[\delta_k\phi_k|\theta_k] - \mathbb{E}[\mathbb{E}[\delta_k|\theta_k]\phi_k] \\ & = & \mathbb{E}[\delta_k\phi_k|\theta_k] - \mathbb{E}[\delta_k|\theta_k]\mathbb{E}[\phi_k] \\ & = & \text{Cov}(\delta_k|\theta_k, \phi_k), \end{array}$$

where $\text{Cov}(\cdot, \cdot)$ is a covariance operator.

Thus,

$$Z_{k+1} \quad = \quad (\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k - \text{Cov}(\delta_k|\theta_k, \phi_k).$$

It is easy to verify that $Z_{k+1}, k \geq 0$ are integrable random variables that satisfy $\mathbb{E}[Z_{k+1}|\mathcal{G}(k)] = 0, \forall k \geq 0$. Also, because ϕ_k, r_k, and ϕ'_k have uniformly bounded second moments, it can be seen that for some constant $c_2 > 0, \forall k \geq 0$,

$$\mathbb{E}[||Z_{k+1}||^2|\mathcal{G}(k)] \leq c_2(1 + ||\theta_k||^2).$$

Consider now the following ODE associated with (5):

$$\begin{array}{rcl} \dot{\theta}(t) & = & \mathrm{Cov}(\delta|\theta(t), \phi) \\ & = & \mathrm{Cov}(r + (\gamma\phi' - \phi)^\top\theta(t), \phi) \\ & = & \mathrm{Cov}(r, \phi) - \mathrm{Cov}(\theta(t)^\top(\phi - \gamma\phi'), \phi) \\ & = & \mathrm{Cov}(r, \phi) - \theta(t)^\top \mathrm{Cov}(\phi - \gamma\phi', \phi) \\ & = & \mathrm{Cov}(r, \phi) - \mathrm{Cov}(\phi - \gamma\phi', \phi)^\top\theta(t) \\ & = & \mathrm{Cov}(r, \phi) - \mathrm{Cov}(\phi, \phi - \gamma\phi')\theta(t) \\ & = & -A\theta(t) + b. \end{array} \tag{A-5}$$

Let $\vec{h}(\theta(t))$ be the driving vector field of the ODE (A-24).

$$\vec{h}(\theta(t)) = -A\theta(t) + b.$$

Consider the cross-covariance matrix,

$$\begin{array}{rcl} A & = & \mathrm{Cov}(\phi, \phi - \gamma\phi') \\ & = & \frac{\mathrm{Cov}(\phi,\phi) + \mathrm{Cov}(\phi-\gamma\phi', \phi-\gamma\phi') - \mathrm{Cov}(\gamma\phi', \gamma\phi')}{2} \\ & = & \frac{\mathrm{Cov}(\phi,\phi) + \mathrm{Cov}(\phi-\gamma\phi', \phi-\gamma\phi') - \gamma^2\mathrm{Cov}(\phi', \phi')}{2} \\ & = & \frac{(1-\gamma^2)\mathrm{Cov}(\phi,\phi) + \mathrm{Cov}(\phi-\gamma\phi', \phi-\gamma\phi')}{2}, \end{array} \tag{A-6}$$

where we eventually used $\mathrm{Cov}(\phi', \phi') = \mathrm{Cov}(\phi, \phi)$ [1]. Note that the covariance matrix $\mathrm{Cov}(\phi, \phi)$ and $\mathrm{Cov}(\phi - \gamma\phi', \phi - \gamma\phi')$ are semi-positive definite. Then, the matrix A is semi-positive definite because A is linearly combined by two positive-weighted semi-positive definite matrice (A-6). Furthermore, A is nonsingular due to the assumption. Hence, the cross-covariance matrix A is positive definite.

Therefore, $\theta^* = A^{-1}b$ can be seen to be the unique globally asymptotically stable equilibrium for ODE (A-24). Let $\vec{h}_\infty(\theta) = \lim_{r\to\infty} \frac{\vec{h}(r\theta)}{r}$. Then $\vec{h}_\infty(\theta) = -A\theta$ is well-defined. Consider now the ODE

$$\dot{\theta}(t) = -A\theta(t). \tag{A-7}$$

The ODE (A-29) has the origin as its unique globally asymptotically stable equilibrium. Thus, the assumption (A1) and (A2) are verified. □

A.2 Proof of Theorem 2

Proof. The proof is similar to that given by (Sutton et al. 2009) for TDC, but it is based on multi-time-scale stochastic approximation.

For the VMTDC algorithm, a new one-step linear TD solution is defined as:

$$0 = \mathbb{E}[(\phi - \gamma\phi' - \mathbb{E}[\phi - \gamma\phi'])\phi^\top]\mathbb{E}[\phi\phi^\top]^{-1}\mathbb{E}[(\delta - \mathbb{E}[\delta])\phi] = \mathbf{A}^\top\mathbf{C}^{-1}(-\mathbf{A}\theta + b).$$

The matrix $\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A}$ is positive definite. Thus, the VMTD's solution is $\theta_{\text{VMTDC}} = \mathbf{A}^{-1}b$.

First, note that recursion (5) and (6) can be rewritten as, respectively,

$$\theta_{k+1} \leftarrow \theta_k + \zeta_k x(k),$$

$$u_{k+1} \leftarrow u_k + \beta_k y(k),$$

where

$$x(k) = \frac{\alpha_k}{\zeta_k}[(\delta_k - \omega_k)\phi_k - \gamma\phi'_k(\phi_k^\top u_k)],$$

$$y(k) = \frac{\zeta_k}{\beta_k}[\delta_k - \omega_k - \phi_k^\top u_k]\phi_k.$$

Recursion (5) can also be rewritten as

$$\theta_{k+1} \leftarrow \theta_k + \beta_k z(k),$$

[1] The covariance matrix $\mathrm{Cov}(\phi', \phi')$ is equal to the covariance matrix $\mathrm{Cov}(\phi, \phi)$ if the initial state is re-reachable or initialized randomly in a Markov chain for on-policy update.

where

$$z(k) = \frac{\alpha_k}{\beta_k}[(\delta_k - \omega_k)\phi_k - \gamma\phi'_k(\phi_k^\top u_k)],$$

Due to the settings of step-size schedule $\alpha_k = o(\zeta_k)$, $\zeta_k = o(\beta_k)$, $x(k) \to 0$, $y(k) \to 0$, $z(k) \to 0$ almost surely as $k \to 0$. That is that the increments in iteration (7) are uniformly larger than those in (6) and the increments in iteration (6) are uniformly larger than those in (5), thus (7) is the fastest recursion, (6) is the second fast recursion and (5) is the slower recursion. Along the fastest time scale, iterations of (5), (6) and (7) are associated to ODEs system as follows:

$$\dot{\theta}(t) = 0, \tag{A-8}$$

$$\dot{u}(t) = 0, \tag{A-9}$$

$$\dot{\omega}(t) = \mathbb{E}[\delta_t | u(t), \theta(t)] - \omega(t). \tag{A-10}$$

Based on the ODE (A-8) and (A-9), both $\theta(t) \equiv \theta$ and $u(t) \equiv u$ when viewed from the fastest timescale. By the Hirsch lemma (Hirsch 1989), it follows that $||\theta_k - \theta|| \to 0$ a.s. as $k \to \infty$ for some θ that depends on the initial condition θ_0 of recursion (5) and $||u_k - u|| \to 0$ a.s. as $k \to \infty$ for some u that depends on the initial condition u_0 of recursion (6). Thus, the ODE pair (A-8)-(refomegavmtdcFastest) can be written as

$$\dot{\omega}(t) = \mathbb{E}[\delta_t | u, \theta] - \omega(t). \tag{A-11}$$

Consider the function $h(\omega) = \mathbb{E}[\delta|\theta, u] - \omega$, i.e., the driving vector field of the ODE (A-11). It is easy to find that the function h is Lipschitz with coefficient -1. Let $h_\infty(\cdot)$ be the function defined by $h_\infty(\omega) = \lim_{r\to\infty} \frac{h(r\omega)}{r}$. Then $h_\infty(\omega) = -\omega$, is well-defined. For (A-11), $\omega^* = \mathbb{E}[\delta|\theta, u]$ is the unique globally asymptotically stable equilibrium. For the ODE

$$\dot{\omega}(t) = h_\infty(\omega(t)) = -\omega(t), \tag{A-12}$$

apply $\vec{V}(\omega) = (-\omega)^\top(-\omega)/2$ as its associated strict Liapunov function. Then, the origin of (A-12) is a globally asymptotically stable equilibrium.

Consider now the recursion (7). Let $M_{k+1} = (\delta_k - \omega_k) - \mathbb{E}[(\delta_k - \omega_k)|\mathcal{F}(k)]$, where $\mathcal{F}(k) = \sigma(\omega_l, u_l, \theta_l, l \le k; \phi_s, \phi'_s, r_s, s < k)$, $k \ge 1$ are the sigma fields generated by $\omega_0, u_0, \theta_0, \omega_{l+1}, u_{l+1}, \theta_{l+1}, \phi_l, \phi'_l$, $0 \le l < k$. It is easy to verify that $M_{k+1}, k \ge 0$ are integrable random variables that satisfy $\mathbb{E}[M_{k+1}|\mathcal{F}(k)] = 0$, $\forall k \ge 0$. Because ϕ_k, r_k, and ϕ'_k have uniformly bounded second moments, it can be seen that for some constant $c_1 > 0$, $\forall k \ge 0$,

$$\mathbb{E}[||M_{k+1}||^2|\mathcal{F}(k)] \le c_1(1 + ||\omega_k||^2 + ||u_k||^2 + ||\theta_k||^2).$$

Now Assumptions (A1) and (A2) of (Borkar and Meyn 2000) are verified. Furthermore, Assumptions (TS) of (Borkar and Meyn 2000) is satisfied by our conditions on the step-size sequences α_k,ζ_k, β_k. Thus, by Theorem 2.2 of (Borkar and Meyn 2000) we obtain that $||\omega_k - \omega^*|| \to 0$ almost surely as $k \to \infty$.

Consider now the second time scale recursion (6). Based on the above analysis, (6) can be rewritten as

$$\dot{\theta}(t) = 0, \tag{A-13}$$

$$\dot{u}(t) = \mathbb{E}[(\delta_t - \mathbb{E}[\delta_t|u(t), \theta(t)])\phi_t|\theta(t)] - \mathbf{C}u(t). \tag{A-14}$$

The ODE (A-13) suggests that $\theta(t) \equiv \theta$ (i.e., a time invariant parameter) when viewed from the second fast timescale. By the Hirsch lemma (Hirsch 1989), it follows that $||\theta_k - \theta|| \to 0$ a.s. as $k \to \infty$ for some θ that depends on the initial condition θ_0 of recursion (5).

Consider now the recursion (6). Let $N_{k+1} = ((\delta_k - \mathbb{E}[\delta_k]) - \phi_k\phi_k^\top u_k) - \mathbb{E}[((\delta_k - \mathbb{E}[\delta_k]) - \phi_k\phi_k^\top u_k)|\mathcal{I}(k)]$, where $\mathcal{I}(k) = \sigma(u_l, \theta_l, l \le k; \phi_s, \phi'_s, r_s, s < k)$, $k \ge 1$ are the sigma fields generated by $u_0, \theta_0, u_{l+1}, \theta_{l+1}, \phi_l, \phi'_l$, $0 \le l < k$. It is easy to verify that $N_{k+1}, k \ge 0$ are integrable random variables that satisfy $\mathbb{E}[N_{k+1}|\mathcal{I}(k)] = 0$, $\forall k \ge 0$. Because ϕ_k, r_k, and ϕ'_k have uniformly bounded second moments, it can be seen that for some constant $c_2 > 0$, $\forall k \ge 0$,

$$\mathbb{E}[||N_{k+1}||^2|\mathcal{I}(k)] \le c_2(1 + ||u_k||^2 + ||\theta_k||^2).$$

Because $\theta(t) \equiv \theta$ from (A-13), the ODE pair (A-13)-(A-14) can be written as

$$\dot{u}(t) = \mathbb{E}[(\delta_t - \mathbb{E}[\delta_t|\theta])\phi_t|\theta] - \mathbf{C}u(t). \tag{A-15}$$

Now consider the function $h(u) = \mathbb{E}[\delta_t - \mathbb{E}[\delta_t|\theta]|\theta] - \mathbf{C}u$, i.e., the driving vector field of the ODE (A-15). For (A-15), $u^* = \mathbf{C}^{-1}\mathbb{E}[(\delta - \mathbb{E}[\delta|\theta])\phi|\theta]$ is the unique globally asymptotically stable equilibrium. Let $h_\infty(u) = -\mathbf{C}u$. For the ODE

$$\dot{u}(t) = h_\infty(u(t)) = -\mathbf{C}u(t), \tag{A-16}$$

the origin of (A-16) is a globally asymptotically stable equilibrium because $\mathbf{C}$ is a positive definite matrix (because it is non-negative definite and nonsingular). Now Assumptions (A1) and (A2) of (Borkar and Meyn 2000) are verified. Furthermore, Assumptions (TS) of (Borkar and Meyn 2000) is satisfied by our conditions on the step-size sequences α_k,ζ_k, β_k. Thus, by Theorem 2.2 of (Borkar and Meyn 2000) we obtain that $||u_k - u^*|| \rightarrow 0$ almost surely as $k \rightarrow \infty$.

Consider now the slower timescale recursion (5). In the light of the above, (5) can be rewritten as

$$\theta_{k+1} \leftarrow \theta_k + \alpha_k(\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k - \alpha_k\gamma\phi'_k(\phi_k^\top \mathbf{C}^{-1}\mathbb{E}[(\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi|\theta_k]). \tag{A-17}$$

Let $\mathcal{G}(k) = \sigma(\theta_l, l \leq k; \phi_s, \phi'_s, r_s, s < k)$, $k \geq 1$ be the sigma fields generated by $\theta_0, \theta_{l+1}, \phi_l, \phi'_l$, $0 \leq l < k$. Let

$$\begin{array}{rcl} Z_{k+1} & = & ((\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k - \gamma\phi'_k\phi_k^\top \mathbf{C}^{-1}\mathbb{E}[(\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi|\theta_k]) \\ & & -\mathbb{E}[((\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k - \gamma\phi'_k\phi_k^\top \mathbf{C}^{-1}\mathbb{E}[(\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi|\theta_k])|\mathcal{G}(k)] \\ & = & ((\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k - \gamma\phi'_k\phi_k^\top \mathbf{C}^{-1}\mathbb{E}[(\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi|\theta_k]) \\ & & -\mathbb{E}[(\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k|\theta_k] - \gamma\mathbb{E}[\phi'\phi^\top]\mathbf{C}^{-1}\mathbb{E}[(\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k|\theta_k]. \end{array}$$

It is easy to see that Z_k, $k \geq 0$ are integrable random variables and $\mathbb{E}[Z_{k+1}|\mathcal{G}(k)] = 0$, $\forall k \geq 0$. Further,

$$\mathbb{E}[||Z_{k+1}||^2|\mathcal{G}(k)] \leq c_3(1 + ||\theta_k||^2), k \geq 0$$

for some constant $c_3 \geq 0$, again beacuse ϕ_k, r_k, and ϕ'_k have uniformly bounded second moments, it can be seen that for some constant.

Consider now the following ODE associated with (5):

$$\dot{\theta}(t) = (\mathbf{I} - \mathbb{E}[\gamma\phi'\phi^\top]\mathbf{C}^{-1})\mathbb{E}[(\delta - \mathbb{E}[\delta|\theta(t)])\phi|\theta(t)]. \tag{A-18}$$

Let

$$\begin{array}{rcl} \vec{h}(\theta(t)) & = & (\mathbf{I} - \mathbb{E}[\gamma\phi'\phi^\top]\mathbf{C}^{-1})\mathbb{E}[(\delta - \mathbb{E}[\delta|\theta(t)])\phi|\theta(t)] \\ & = & (\mathbf{C} - \mathbb{E}[\gamma\phi'\phi^\top])\mathbf{C}^{-1}\mathbb{E}[(\delta - \mathbb{E}[\delta|\theta(t)])\phi|\theta(t)] \\ & = & (\mathbb{E}[\phi\phi^\top] - \mathbb{E}[\gamma\phi'\phi^\top])\mathbf{C}^{-1}\mathbb{E}[(\delta - \mathbb{E}[\delta|\theta(t)])\phi|\theta(t)] \\ & = & \mathbf{A}^\top \mathbf{C}^{-1}(-\mathbf{A}\theta(t) + b), \end{array}$$

because $\mathbb{E}[(\delta - \mathbb{E}[\delta|\theta(t)])\phi|\theta(t)] = -\mathbf{A}\theta(t) + b$, where $\mathbf{A} = \mathrm{Cov}(\phi, \phi - \gamma\phi')$, $b = \mathrm{Cov}(r, \phi)$, and $\mathbf{C} = \mathbb{E}[\phi\phi^\top]$

Therefore, $\theta^* = \mathbf{A}^{-1}b$ can be seen to be the unique globally asymptotically stable equilibrium for ODE (A-18). Let $\vec{h}_\infty(\theta) = \lim_{r\rightarrow\infty} \frac{\vec{h}(r\theta)}{r}$. Then $\vec{h}_\infty(\theta) = -\mathbf{A}^\top \mathbf{C}^{-1}\mathbf{A}\theta$ is well-defined. Consider now the ODE

$$\dot{\theta}(t) = -\mathbf{A}^\top \mathbf{C}^{-1}\mathbf{A}\theta(t). \tag{A-19}$$

Because $\mathbf{C}^{-1}$ is positive definite and $\mathbf{A}$ has full rank (as it is nonsingular by assumption), the matrix $\mathbf{A}^\top \mathbf{C}^{-1}\mathbf{A}$ is also positive definite. The ODE (A-19) has the origin as its unique globally asymptotically stable equilibrium. Thus, the assumption (A1) and (A2) are verified.

The proof is given above. In the fastest time scale, the parameter w converges to $\mathbb{E}[\delta|u_k, \theta_k]$. In the second fast time scale, the parameter u converges to $\mathbf{C}^{-1}\mathbb{E}[(\delta - \mathbb{E}[\delta|\theta_k])\phi|\theta_k]$. In the slower time scale, the parameter θ converges to $\mathbf{A}^{-1}b$. □

A.3 Proof of Theorem 2

Proof. The proof of VMETD's convergence is also based on Borkar's Theorem for general stochastic approximation recursions with two time scales (Borkar 1997).

The VMTD's solution is $\theta_{\text{VMETD}} = \mathbf{A}_{\text{VMETD}}^{-1} b_{\text{VMETD}}$.

First, note that recursion (19) can be rewritten as

$$\theta_{k+1} \leftarrow \theta_k + \beta_k\xi(k),$$

where

$$\xi(k) = \frac{\alpha_k}{\beta_k}(F_k\rho_k\delta_k - \omega_{k+1})\phi_k$$

Due to the settings of step-size schedule $\alpha_k = o(\beta_k)$, $\xi(k) \rightarrow 0$ almost surely as $k \rightarrow \infty$. That is the increments in iteration (13) are uniformly larger than those in (12), thus (13) is the faster recursion. Along the faster time scale, iterations of (12) and (13) are associated to ODEs system as follows:

$$\dot{\theta}(t) = 0, \tag{A-20}$$

$$\dot{\omega}(t) = \mathbb{E}_\mu[F_t\rho_t\delta_t|\theta(t)] - \omega(t). \tag{A-21}$$

Based on the ODE (A-20), $\theta(t) \equiv \theta$ when viewed from the faster timescale. By the Hirsch lemma (Hirsch 1989), it follows that $||\theta_k - \theta|| \rightarrow 0$ a.s. as $k \rightarrow \infty$ for some θ that depends on the initial condition θ_0 of recursion (12). Thus, the ODE pair (A-20)-(A-21) can be written as

$$\dot{\omega}(t) = \mathbb{E}_\mu[F_t\rho_t\delta_t|\theta] - \omega(t). \tag{A-22}$$

Consider the function $h(\omega) = \mathbb{E}_\mu[F\rho\delta|\theta] - \omega$, i.e., the driving vector field of the ODE (A-22). It is easy to find that the function h is Lipschitz with coefficient -1. Let $h_\infty(\cdot)$ be the function defined by $h_\infty(\omega) = \lim_{x\to\infty} \frac{h(x\omega)}{x}$. Then $h_\infty(\omega) = -\omega$, is well-defined. For (A-22), $\omega^* = \mathbb{E}_\mu[F\rho\delta|\theta]$ is the unique globally asymptotically stable equilibrium. For the ODE

$$\dot{\omega}(t) = h_\infty(\omega(t)) = -\omega(t), \tag{A-23}$$

apply $\vec{V}(\omega) = (-\omega)^\top(-\omega)/2$ as its associated strict Liapunov function. Then, the origin of (A-23) is a globally asymptotically stable equilibrium.

Consider now the recursion (13). Let $M_{k+1} = (F_k\rho_k\delta_k - \omega_k) - \mathbb{E}_\mu[(F_k\rho_k\delta_k - \omega_k)|\mathcal{F}(k)]$, where $\mathcal{F}(k) = \sigma(\omega_l, \theta_l, l \le k; \phi_s, \phi'_s, r_s, s < k)$, $k \ge 1$ are the sigma fields generated by $\omega_0, \theta_0, \omega_{l+1}, \theta_{l+1}, \phi_l, \phi'_l$, $0 \le l < k$. It is easy to verify that $M_{k+1}, k \ge 0$ are integrable random variables that satisfy $\mathbb{E}[M_{k+1}|\mathcal{F}(k)] = 0, \forall k \ge 0$. Because ϕ_k, r_k, and ϕ'_k have uniformly bounded second moments, it can be seen that for some constant $c_1 > 0, \forall k \ge 0$,

$$\mathbb{E}[||M_{k+1}||^2|\mathcal{F}(k)] \le c_1(1 + ||\omega_k||^2 + ||\theta_k||^2).$$

Now Assumptions (A1) and (A2) of (Borkar and Meyn 2000) are verified. Furthermore, Assumptions (TS) of (Borkar and Meyn 2000) is satisfied by our conditions on the step-size sequences α_k, β_k. Thus, by Theorem 2.2 of (Borkar and Meyn 2000) we obtain that $||\omega_k - \omega^*|| \to 0$ almost surely as $k \to \infty$.

Consider now the slower time scale recursion (12). Based on the above analysis, (12) can be rewritten as

$$\begin{aligned}
\theta_{k+1} &\leftarrow \theta_k + \alpha_k(F_k\rho_k\delta_k - \omega_k)\phi_k - \alpha_k\omega_{k+1}\phi_k \\
&= \theta_k + \alpha_k(F_k\rho_k\delta_k - \mathbb{E}_\mu[F_k\rho_k\delta_k|\theta_k])\phi_k \\
&= \theta_k + \alpha_k F_k\rho_k(R_{k+1} + \gamma\theta_k^\top\phi_{k+1} - \theta_k^\top\phi_k)\phi_k - \alpha_k\mathbb{E}_\mu[F_k\rho_k\delta_k]\phi_k \\
&= \theta_k + \alpha_k\{\underbrace{(F_k\rho_k R_{k+1} - \mathbb{E}_\mu[F_k\rho_k R_{k+1}])\phi_k}_{b_{\text{VMETD},k}} - \underbrace{(F_k\rho_k\phi_k(\phi_k - \gamma\phi_{k+1})^\top - \phi_k\mathbb{E}_\mu[F_k\rho_k(\phi_k - \gamma\phi_{k+1})]^\top)}_{\mathbf{A}_{\text{VMETD},k}}\theta_k\}
\end{aligned}$$

Let $\mathcal{G}(k) = \sigma(\theta_l, l \le k; \phi_s, \phi'_s, r_s, s < k)$, $k \ge 1$ be the sigma fields generated by $\theta_0, \theta_{l+1}, \phi_l, \phi'_l$, $0 \le l < k$. Let $Z_{k+1} = Y_k - \mathbb{E}[Y_k|\mathcal{G}(k)]$, where

$$Y_k = (F_k\rho_k\delta_k - \mathbb{E}_\mu[F_k\rho_k\delta_k|\theta_k])\phi_k.$$

Consequently,

$$\begin{aligned}
\mathbb{E}_\mu[Y_k|\mathcal{G}(k)] &= \mathbb{E}_\mu[(F_k\rho_k\delta_k - \mathbb{E}_\mu[F_k\rho_k\delta_k|\theta_k])\phi_k|\mathcal{G}(k)] \\
&= \mathbb{E}_\mu[F_k\rho_k\delta_k\phi_k|\theta_k] - \mathbb{E}_\mu[\mathbb{E}_\mu[F_k\rho_k\delta_k|\theta_k]\phi_k] \\
&= \mathbb{E}_\mu[F_k\rho_k\delta_k\phi_k|\theta_k] - \mathbb{E}_\mu[F_k\rho_k\delta_k|\theta_k]\mathbb{E}_\mu[\phi_k] \\
&= \mathrm{Cov}(F_k\rho_k\delta_k|\theta_k, \phi_k),
\end{aligned}$$

where $\mathrm{Cov}(\cdot,\cdot)$ is a covariance operator.

Thus,

$$Z_{k+1} = (F_k\rho_k\delta_k - \mathbb{E}[\delta_k|\theta_k])\phi_k - \mathrm{Cov}(F_k\rho_k\delta_k|\theta_k, \phi_k).$$

It is easy to verify that $Z_{k+1}, k \ge 0$ are integrable random variables that satisfy $\mathbb{E}[Z_{k+1}|\mathcal{G}(k)] = 0, \forall k \ge 0$. Also, because ϕ_k, r_k, and ϕ'_k have uniformly bounded second moments, it can be seen that for some constant $c_2 > 0, \forall k \ge 0$,

$$\mathbb{E}[||Z_{k+1}||^2|\mathcal{G}(k)] \le c_2(1 + ||\theta_k||^2).$$

Consider now the following ODE associated with (12):

$$\dot{\theta}(t) = -\mathbf{A}_{\text{VMETD}}\theta(t) + b_{\text{VMETD}}. \tag{A-24}$$

$$\begin{aligned}
\mathbf{A}_{\text{VMETD}} &= \lim_{k\to\infty}\mathbb{E}[\mathbf{A}_{\text{VMETD},k}] \\
&= \lim_{k\to\infty}\mathbb{E}_\mu[F_k\rho_k\phi_k(\phi_k - \gamma\phi_{k+1})^\top] - \lim_{k\to\infty}\mathbb{E}_\mu[\phi_k]\mathbb{E}_\mu[F_k\rho_k(\phi_k - \gamma\phi_{k+1})]^\top \\
&= \lim_{k\to\infty}\mathbb{E}_\mu[\phi_k F_k\rho_k(\phi_k - \gamma\phi_{k+1})^\top] - \lim_{k\to\infty}\mathbb{E}_\mu[\phi_k]\mathbb{E}_\mu[F_k\rho_k(\phi_k - \gamma\phi_{k+1})]^\top \\
&= \lim_{k\to\infty}\mathbb{E}_\mu[\phi_k F_k\rho_k(\phi_k - \gamma\phi_{k+1})^\top] - \lim_{k\to\infty}\mathbb{E}_\mu[\phi_k]\lim_{k\to\infty}\mathbb{E}_\mu[F_k\rho_k(\phi_k - \gamma\phi_{k+1})]^\top \\
&= \sum_s f(s)\phi(s)(\phi(s) - \gamma\sum_{s'}[\mathbf{P}_\pi]_{ss'}\phi(s'))^\top - \sum_s d_\mu(s)\phi(s) * \sum_s f(s)(\phi(s) - \gamma\sum_{s'}[\mathbf{P}_\pi]_{ss'}\phi(s'))^\top \qquad \text{(A-25)} \\
&= \Phi^\top\mathbf{F}(\mathbf{I} - \gamma\mathbf{P}_\pi)\Phi - \Phi^\top\mathbf{d}_\mu\mathbf{f}^\top(\mathbf{I} - \gamma\mathbf{P}_\mu)\Phi \\
&= \Phi^\top(\mathbf{F} - \mathbf{d}_\mu\mathbf{f}^\top)(\mathbf{I} - \gamma\mathbf{P}_\pi)\Phi \\
&= \Phi^\top(\mathbf{F}(\mathbf{I} - \gamma\mathbf{P}_\pi) - \mathbf{d}_\mu\mathbf{f}^\top(\mathbf{I} - \gamma\mathbf{P}_\pi))\Phi \\
&= \Phi^\top(\mathbf{F}(\mathbf{I} - \gamma\mathbf{P}_\pi) - \mathbf{d}_\mu\mathbf{d}_\mu^\top)\Phi
\end{aligned}$$

$$\begin{aligned}
b_{\text{VMETD}} &= \lim_{k\to\infty} \mathbb{E}[b_{\text{VMETD},k}] \\
&= \lim_{k\to\infty} \mathbb{E}_\mu[F_k\rho_k R_{k+1}\phi_k] - \lim_{k\to\infty} \mathbb{E}_\mu[\phi_k]\mathbb{E}_\mu[F_k\rho_k R_{k+1}] \\
&= \lim_{k\to\infty} \mathbb{E}_\mu[\phi_k F_k\rho_k R_{k+1}] - \lim_{k\to\infty} \mathbb{E}_\mu[\phi_k]\mathbb{E}_\mu[\phi_k]\mathbb{E}_\mu[F_k\rho_k R_{k+1}] \\
&= \lim_{k\to\infty} \mathbb{E}_\mu[\phi_k F_k\rho_k R_{k+1}] - \lim_{k\to\infty} \mathbb{E}_\mu[\phi_k] \lim_{k\to\infty} \mathbb{E}_\mu[F_k\rho_k R_{k+1}] \\
&= \sum_s f(s)\phi(s)r_\pi - \sum_s d_\mu(s)\phi(s) * \sum_s f(s)r_\pi \\
&= \Phi^\top(\mathbf{F} - \mathbf{d}_\mu\mathbf{f}^\top)\mathbf{r}_\pi
\end{aligned} \tag{A-26}$$

Let $\vec{h}(\theta(t))$ be the driving vector field of the ODE (A-24).

$$\vec{h}(\theta(t)) = -\mathbf{A}_{\text{VMETD}}\theta(t) + b_{\text{VMETD}}.$$

An $\Phi^\top \mathrm{X}\Phi$ matrix of this form will be positive definite whenever the matrix X is positive definite. Any matrix X is positive definite if and only if the symmetric matrix $\mathrm{S} = \mathrm{X} + \mathrm{X}^\top$ is positive definite. Any symmetric real matrix S is positive definite if the absolute values of its diagonal entries are greater than the sum of the absolute values of the corresponding off-diagonal entries(Sutton, Mahmood, and White 2016).

$$\begin{aligned}
(\mathbf{F}(\mathbf{I} - \gamma\mathbf{P}_\pi) - \mathbf{d}_\mu\mathbf{d}_\mu^\top)\mathbf{1} &= \mathbf{F}(\mathbf{I} - \gamma\mathbf{P}_\pi)\mathbf{1} - \mathbf{d}_\mu\mathbf{d}_\mu^\top\mathbf{1} \\
&= \mathbf{F}(\mathbf{1} - \gamma\mathbf{P}_\pi\mathbf{1}) - \mathbf{d}_\mu\mathbf{d}_\mu^\top\mathbf{1} \\
&= (1-\gamma)\mathbf{F}\mathbf{1} - \mathbf{d}_\mu\mathbf{d}_\mu^\top\mathbf{1} \\
&= (1-\gamma)\mathbf{f} - \mathbf{d}_\mu\mathbf{d}_\mu^\top\mathbf{1} \\
&= (1-\gamma)\mathbf{f} - \mathbf{d}_\mu \\
&= (1-\gamma)(\mathbf{I} - \gamma\mathbf{P}_\pi^\top)^{-1}\mathbf{d}_\mu - \mathbf{d}_\mu \\
&= (1-\gamma)[(\mathbf{I} - \gamma\mathbf{P}_\pi^\top)^{-1} - \mathbf{I}]\mathbf{d}_\mu \\
&= (1-\gamma)[\sum_{t=0}^{\infty}(\gamma\mathbf{P}_\pi^\top)^t - \mathbf{I}]\mathbf{d}_\mu \\
&= (1-\gamma)[\sum_{t=1}^{\infty}(\gamma\mathbf{P}_\pi^\top)^t]\mathbf{d}_\mu > 0
\end{aligned} \tag{A-27}$$

$$\begin{aligned}
\mathbf{1}^\top(\mathbf{F}(\mathbf{I} - \gamma\mathbf{P}_\pi) - \mathbf{d}_\mu\mathbf{d}_\mu^\top) &= \mathbf{1}^\top\mathbf{F}(\mathbf{I} - \gamma\mathbf{P}_\pi) - \mathbf{1}^\top\mathbf{d}_\mu\mathbf{d}_\mu^\top \\
&= \mathbf{d}_\mu^\top - \mathbf{1}^\top\mathbf{d}_\mu\mathbf{d}_\mu^\top \\
&= \mathbf{d}_\mu^\top - \mathbf{d}_\mu^\top \\
&= 0
\end{aligned} \tag{A-28}$$

(A-27) and (A-28) show that the matrix $\mathbf{F}(\mathbf{I} - \gamma\mathbf{P}_\pi) - \mathbf{d}_\mu\mathbf{d}_\mu^\top$ of diagonal entries are positive and its off-diagonal entries are negative. So its each row sum plus the corresponding column sum is positive. So $\mathbf{A}_{\text{VMETD}}$ is positive definite.

Therefore, $\theta^* = \mathbf{A}_{\text{VMETD}}^{-1}b_{\text{VMETD}}$ can be seen to be the unique globally asymptotically stable equilibrium for ODE (A-24). Let $\vec{h}_\infty(\theta) = \lim_{r\to\infty}\frac{\vec{h}(r\theta)}{r}$. Then $\vec{h}_\infty(\theta) = -\mathbf{A}_{\text{VMETD}}\theta$ is well-defined. Consider now the ODE

$$\dot{\theta}(t) = -\mathbf{A}_{\text{VMETD}}\theta(t). \tag{A-29}$$

The ODE (A-29) has the origin as its unique globally asymptotically stable equilibrium. Thus, the assumption (A1) and (A2) are verified. □

B Experimental details

2-state version of Baird's off-policy counterexample: All learning rates follow linear learning rate decay. For TD algorithm, $\frac{\alpha_k}{\omega_k} = 4$ and $\alpha_0 = 0.1$. For TDC algorithm, $\frac{\alpha_k}{\zeta_k} = 5$ and $\alpha_0 = 0.1$. For VMTDC algorithm, $\frac{\alpha_k}{\zeta_k} = 5$, $\frac{\alpha_k}{\omega_k} = 4$,and $\alpha_0 = 0.1$. For VMTD algorithm, $\frac{\alpha_k}{\omega_k} = 4$ and $\alpha_0 = 0.1$.

2-state version of Baird's off-policy counterexample: All learning rates follow linear learning rate decay. For TD algorithm, $\frac{\alpha_k}{\omega_k} = 4$ and $\alpha_0 = 0.1$. For TDC algorithm, $\frac{\alpha_k}{\zeta_k} = 5$ and $\alpha_0 = 0.1$.For ETD algorithm, $\alpha_0 = 0.1$. For VMTDC algorithm, $\frac{\alpha_k}{\zeta_k} = 5$, $\frac{\alpha_k}{\omega_k} = 4$,and $\alpha_0 = 0.1$.For VMETD algorithm, $\frac{\alpha_k}{\omega_k} = 4$ and $\alpha_0 = 0.1$. For VMTD algorithm, $\frac{\alpha_k}{\omega_k} = 4$ and $\alpha_0 = 0.1$.

For all policy evaluation experiments, each experiment is independently run 100 times.

For the four control experiments: The learning rates for each algorithm in all experiments are shown in Table 1. For all control experiments, each experiment is independently run 50 times.

Table 1: Learning rates (lr) of four control experiments.

algorithms(lr) \ envs	Maze	Cliff walking	Mountain Car	Acrobot
Sarsa(α)	0.1	0.1	0.1	0.1
GQ(α, ζ)	$0.1, 0.003$	$0.1, 0.004$	$0.1, 0.01$	$0.1, 0.01$
EQ(α)	0.006	0.005	0.001	0.0005
VMSarsa(α, β)	$0.1, 0.001$	$0.1,$ 1e-4	$0.1,$ 1e-4	$0.1,$ 1e-4
VMGQ(α, ζ, β)	$0.1, 0.001, 0.001$	$0.1, 0.005,$ 1e-4	$0.1,$ 5e-4, 1e-4	$0.1,$ 5e-4, 1e-4
VMEQ(α, β)	$0.001, 0.0005$	$0.005, 0.0001$	$0.001, 0.0001$	$0.0005, 0.0001$
Q-learning(α)	0.1	0.1	0.1	0.1
VMQ(α, β)	$0.1, 0.001$	$0.1,$ 1e-4	$0.1,$ 1e-4	$0.1,$ 1e-4

References

Borkar, V. S. 1997. Stochastic approximation with two time scales. *Syst. & Control Letters*, 29(5): 291–294.

Borkar, V. S.; and Meyn, S. P. 2000. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2): 447–469.

Hirsch, M. W. 1989. Convergent activation dynamics in continuous time networks. *Neural Netw.*, 2(5): 331–349.

Sutton, R.; Maei, H.; Precup, D.; Bhatnagar, S.; Silver, D.; Szepesvári, C.; and Wiewiora, E. 2009. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proc. 26th Int. Conf. Mach. Learn.*, 993–1000.

Sutton, R. S.; Mahmood, A. R.; and White, M. 2016. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17(1): 2603–2631.