Time-Variant Variational Transfer for Value Functions: Supplementary Material

Giuseppe Canonaco [*1] **Andrea Soprani** [*1] **Matteo Giuliani**[1] **Andrea Castelletti**[1] **Manuel Roveri**[1] **Marcello Restelli**[1]

[1]Department of Electronics Information and Bioengineering, Politecnico di Milano, Milan, Italy

A PROOF OF THEOREM 3.6

Definition A.1. For a spatial kernel K_S: $\mu_l(K_S) = \int \theta^l K_S(\theta) d\theta$

Definition A.2. For a temporal kernel K_T: $a_l(-\rho) = \int_{-\rho}^{1} t^l K_T(t) dt$

Lemma A.1 (Estimator consistency on the right boundary)**.** *Let $t \in B_r = \{\tau : 1 - \lambda \leq \tau \leq 1\}$ then under assumptions of Theorem 3.6:*

$$\mathbb{E}[\hat{p}(\theta, t)|\mathcal{M}] = p(\theta, t) + O(\lambda) + O(tr(H)),$$

where $\mathcal{M}$ represents all the discrete random variables M_i for $i = 1 \dots n$.

Proof.

$$\mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}] =$$

$$= \frac{1}{\bar{N}\lambda|H|^{\frac{1}{2}}a_0(-\rho)} \sum_{i=1}^{n} \int K_T\left(\frac{t-\tau}{\lambda}\right) \sum_{j=1}^{M_i} \int_{-\infty}^{+\infty} K_S\left(H^{-\frac{1}{2}}(\theta - x)\right) p(x,\tau) dx d\tau \tag{1}$$

$$= \frac{1}{\bar{N}\lambda\cancel{|H|^{\frac{1}{2}}}a_0(-\rho)} \sum_{i=1}^{n} \int K_T\left(\frac{t-\tau}{\lambda}\right) \sum_{j=1}^{M_i} \int_{+\infty}^{-\infty} -K_S(y) p(\theta - H^{\frac{1}{2}}y, \tau) \cancel{|H|^{\frac{1}{2}}} dy d\tau \tag{2}$$

$$= \frac{1}{\bar{N}\lambda a_0(-\rho)} \sum_{i=1}^{n} \int K_T\left(\frac{t-\tau}{\lambda}\right) \sum_{j=1}^{M_i} \int_{-\infty}^{+\infty} K_S(y) \Big(p(\theta,\tau) - (H^{\frac{1}{2}}y)^T \nabla^S p(\theta,\tau) +$$
$$\frac{1}{2}(H^{\frac{1}{2}}y)^T \mathcal{H}^S p(\theta,\tau)(H^{\frac{1}{2}}y) + o(tr(H)) \Big) dy d\tau \tag{3}$$

$$= \frac{1}{\bar{N}\lambda a_0(-\rho)} \sum_{i=1}^{n} \int K_T\left(\frac{t-\tau}{\lambda}\right) M_i \Big(\int_{-\infty}^{+\infty} K_S(y) p(\theta,\tau) dy$$
$$\cancel{- \int_{-\infty}^{+\infty} K_S(y)(H^{\frac{1}{2}}y)^T \nabla^S p(\theta,\tau) dy} +$$
$$\int_{-\infty}^{+\infty} \frac{1}{2} K_S(y)(H^{\frac{1}{2}}y)^T \mathcal{H}^S p(\theta,\tau)(H^{\frac{1}{2}}y) dy + o(tr(H)) \Big) d\tau \tag{4}$$

$$= \frac{1}{\bar{N}\lambda a_0(-\rho)} \sum_{i=1}^{n} \int K_T\left(\frac{t-\tau}{\lambda}\right) M_i \Big(p(\theta,\tau) +$$

[*]equal contribution

$$\frac{1}{2}\mu_2(K_S)tr(H\mathcal{H}^s p(\theta,\tau)) + o(tr(H))\Big)d\tau \tag{5}$$

$$= \frac{1}{\lambda a_0(-\rho)}\int_{\frac{t-1}{\lambda}}^{\frac{t}{\lambda}} K_T\left(\frac{t-\tau}{\lambda}\right)\Big(p(\theta,\tau) + O(tr(H))\Big)d\tau \tag{6}$$

$$= \frac{1}{\lambda a_0(-\rho)}\left(\int_{-\rho}^{1} K_T\left(\frac{t-\tau}{\lambda}\right)p(\theta,\tau)d\tau + O(tr(H))\int_{-\rho}^{1} K_T\left(\frac{t-\tau}{\lambda}\right)d\tau\right) \tag{7}$$

$$= \frac{\lambda}{\lambda a_0(-\rho)}\left(-\int_{1}^{-\rho} K_T(v)p(\theta,t-\lambda v)dv - O(tr(H))\int_{1}^{-\rho} K_T(v)dv\right) \tag{8}$$

$$= \frac{1}{a_0(-\rho)}\Bigg(\int_{-\rho}^{1} K_T(v)\Big(p(\theta,t) - \lambda v p'(\theta,t) + \frac{1}{2}\lambda^2 v^2 p''(\theta,t) + o(\lambda^2)\Big)dv + O(tr(H))\Bigg) \tag{9}$$

$$= p(\theta,t) - \lambda p'(\theta,t)\frac{a_1(-\rho)}{a_0(-\rho)} + O(\lambda^2) + O(tr(H)) \tag{10}$$

$$= p(\theta,t) + O(\lambda) + O(tr(H)), \tag{11}$$

where in (2) we performed a change of variable, $y = H^{-\frac{1}{2}}(\theta - x)$, in (3) we used the following Taylor expansion:

$$p(\theta - H^{\frac{1}{2}}y,\tau) = p(\theta,\tau) - (H^{\frac{1}{2}}y)^T\nabla^S p(\theta,\tau) + \frac{1}{2}(H^{\frac{1}{2}}y)^T\mathcal{H}^S p(\theta,\tau)(H^{\frac{1}{2}}y) + o(tr(H)),$$

in (4) we used Assumption 3.4,in (5) we used Definition A.1, in (6) we used $\frac{t-\tau}{\lambda} \in [\frac{t-1}{\lambda}, \frac{t}{\lambda})$, in (7) we set $t = 1 - \rho\lambda$, which implies $\frac{t-\tau}{\lambda} \in [-\rho, \frac{1}{\lambda} - \rho)$, then we used the support of K_T (assumed to be $[-1,1]$ without loss of generality) since $\lambda \to 0$. Finally, in (8) we used a change of variable, $\frac{t-\tau}{\lambda} = v$, and in (9) we used the following Taylor expansion:

$$p(\theta,t-\lambda v) = p(\theta,t) - \lambda v p'(\theta,t) + \frac{1}{2}\lambda^2 v^2 p''(\theta,v) + o(\lambda^2).$$

□

Notice that we reported the consistency proof only on the right boundary because is the one we use in the context of our algorithm. The above procedure can be easily adjusted to prove consistency of the estimator on the left boundary getting the same convergence rate. Moreover, analogously, we can obtain consistency away from the two boundaries with a convergence rate squared w.r.t. λ.

Definition A.3. For a spatial kernel K_S: $R(K_S) = \int K_S^2(\theta)d\theta$

Definition A.4. For a temporal kernel K_T: $b_{K_T}(-\rho) = \int_{-\rho}^{1} K_T^2(t)dt$

Lemma A.2 (Variance of the estimator on the right boundary)**.** *Let $t \in B_r = \{\tau : 1 - \lambda \leq \tau \leq 1\}$ then under assumptions of Theorem 3.6:*

$$\mathbb{V}ar[\hat{p}(\theta,t)|\mathcal{M}] \leq \frac{C_1}{\bar{N}|H|^{\frac{1}{2}}\lambda},$$

where $\mathcal{M}$ represents all the discrete random variables M_i for $i = 1 \dots n$.

Proof.

$$\mathbb{V}ar[\hat{p}(\theta,t)|\mathcal{M}] = \frac{1}{\bar{N}a_0^2(-\rho)}\mathbb{V}ar\left[\frac{1}{|H|^{\frac{1}{2}}\lambda}K_T\left(\frac{t-t_i}{\lambda}\right)K_S\left(H^{-\frac{1}{2}}(\theta - x_{ij})\right)\right] \tag{12}$$

$$= \frac{1}{\bar{N}a_0^2(-\rho)}\Bigg(\mathbb{E}\left[\frac{1}{|H|\lambda^2}K_T^2\left(\frac{t-t_i}{\lambda}\right)K_S^2\left(H^{-\frac{1}{2}}(\theta - x_{ij})\right)\right] - \mathbb{E}^2\left[\frac{1}{|H|^{\frac{1}{2}}\lambda}K_T\left(\frac{t-t_i}{\lambda}\right)K_S\left(H^{-\frac{1}{2}}(\theta - x_{ij})\right)\right]\Bigg) \tag{13}$$

$$= \frac{1}{\bar{N}a_0^2(-\rho)}\Bigg(\int \frac{1}{|H|\lambda^2}K_T^2\left(\frac{t-\tau}{\lambda}\right)\int_{+\infty}^{-\infty} -|H|^{\frac{1}{2}}K_S^2(y)p(\theta - H^{\frac{1}{2}}y,\tau)dyd\tau -$$
$$\left(\int \frac{1}{|H|^{\frac{1}{2}}\lambda}K_T\left(\frac{t-\tau}{\lambda}\right)\int_{+\infty}^{-\infty} -|H|^{\frac{1}{2}}K_S(y)p(\theta - H^{\frac{1}{2}}y,\tau)dyd\tau\right)^2\Bigg) \tag{14}$$

$$= \frac{1}{\bar{N}a_0^2(-\rho)}\Bigg(\int \frac{1}{|H|^{\frac{1}{2}}\lambda^2}K_T^2\left(\frac{t-\tau}{\lambda}\right)\int_{-\infty}^{+\infty} K_S^2(y)\left(p(\theta,\tau)+o(1)\right)dyd\tau -$$
$$\left(\int \frac{1}{\lambda}K_T\left(\frac{t-\tau}{\lambda}\right)\int_{-\infty}^{+\infty} K_S(y)\left(p(\theta,\tau)+o(1)\right)dyd\tau\right)^2\Bigg) \tag{15}$$

$$= \frac{1}{\bar{N}a_0^2(-\rho)}\Bigg(\int \frac{1}{|H|^{\frac{1}{2}}\lambda^2}K_T^2\left(\frac{t-\tau}{\lambda}\right)\left(p(\theta,\tau)+o(1)\right)R(K_S)d\tau -$$
$$\left(\int \frac{1}{\lambda}K_T\left(\frac{t-\tau}{\lambda}\right)\left(p(\theta,\tau)+o(1)\right)d\tau\right)^2\Bigg) \tag{16}$$

$$= \frac{1}{\bar{N}a_0^2(-\rho)}\Bigg(\int_{-\rho}^{1} \frac{1}{|H|^{\frac{1}{2}}\lambda^2}K_T^2\left(\frac{t-\tau}{\lambda}\right)\left(p(\theta,\tau)+o(1)\right)R(K_S)d\tau -$$
$$\left(\int_{-\rho}^{1} \frac{1}{\lambda}K_T\left(\frac{t-\tau}{\lambda}\right)\left(p(\theta,\tau)+o(1)\right)d\tau\right)^2\Bigg) \tag{17}$$

$$= \frac{1}{\bar{N}a_0^2(-\rho)}\Bigg(\int_{1}^{-\rho} -\frac{1}{|H|^{\frac{1}{2}}\lambda}K_T^2(v)\left(p(\theta,t-\lambda v)+o(1)\right)R(K_S)dv -$$
$$\left(\int_{1}^{-\rho} -K_T(v)\left(p(\theta,t-\lambda v)+o(1)\right)dv\right)^2\Bigg) \tag{18}$$

$$= \frac{1}{\bar{N}a_0^2(-\rho)}\Bigg(\int_{-\rho}^{1} \frac{1}{|H|^{\frac{1}{2}}\lambda}K_T^2(v)\left(p(\theta,t)+o(1)\right)R(K_S)dv -$$
$$\left(\int_{-\rho}^{1} K_T(v)\left(p(\theta,t)+o(1)\right)dv\right)^2\Bigg) \tag{19}$$

$$= \frac{1}{\bar{N}a_0^2(-\rho)}\left(\frac{p(\theta,t)+o(1)}{|H|^{\frac{1}{2}}\lambda}R(K_S)b_{K_T}(-\rho) - \left(a_0(-\rho)(p(\theta,t)+o(1))\right)^2\right) \tag{20}$$

$$= \frac{p(\theta,t)R(K_S)b_{K_T}(-\rho)}{\bar{N}|H|^{\frac{1}{2}}\lambda a_0^2(-\rho)} + O\left(\frac{1}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right) \tag{21}$$

$$= O\left(\frac{1}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right) \quad \rightarrow \quad \exists\, C_1 : \mathbb{V}ar[\hat{p}(\theta,t)|\mathcal{M}] \leq \frac{C_1}{\bar{N}|H|^{\frac{1}{2}}\lambda}, \tag{22}$$

where in (14) we performed a change of variable, $y = H^{-\frac{1}{2}}(\theta - x)$, in (15) we used the following Taylor expansion:

$$p(\theta - H^{\frac{1}{2}}y,\tau) = p(\theta,\tau) + o(1),$$

in (16) we used Definition A.3, in (17) we considered the fact that $t \in B_r$ as we have done in (6) and (7) of the proof of A.1, in (18) we performed a change of variable, $\frac{t-\tau}{\lambda} = v$, in (19) we used the following Taylor expansion:

$$p(\theta,t-\lambda v) = p(\theta,t) + o(1),$$

whereas in (20) we have used Definition A.4. Finally, in (22) we have used the fact that $p(\theta,t)$ has bounded derivatives and is a pdf, therefore it has finite supremum. □

Lemma A.3 (Bound on the absolute values)**.** *Let $t \in B_r = \{\tau : 1-\lambda \leq \tau \leq 1\}$ then under assumptions of Theorem 3.6: $\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]$ is the sum of $\bar{N}$ independent random variables, denoted as v_i, with zero mean and absolute values bounded by $\frac{C_2}{\bar{N}|H|^{\frac{1}{2}}\lambda}$. $\mathcal{M}$ represents all the discrete random variables M_i for $i = 1 \dots n$.*

Proof.

$$|v_i| = \left|\frac{1}{\bar{N}\lambda|H|^{\frac{1}{2}}a_0(-\rho)}K_T\left(\frac{t-t_i}{\lambda}\right)K_S(H^{-\frac{1}{2}}(\theta - x_{ij})) -\right.$$

$$\frac{p(\theta,t)+O(\lambda)+O(tr(H))}{\bar{N}}\Big| \tag{23}$$

$$\leq \left|\frac{M_T M_S}{\bar{N}\lambda|H|^{\frac{1}{2}}a_0(-\rho)} - \frac{p(\theta,t)+O(\lambda)+O(tr(H))}{\bar{N}}\right| \tag{24}$$

$$= O\left(\frac{1}{\bar{N}\lambda|H|^{\frac{1}{2}}}\right) \rightarrow \exists C_2 :\ |v_i| \leq \frac{C_2}{\bar{N}|H|^{\frac{1}{2}}\lambda}, \tag{25}$$

where in (23) we used *lemma* A.1 and in (24) we used the fact that K_T has a compact support on $\mathbb{R}$ and K_S has a supremum. □

Now the proof of *Theorem* 3.6 can follow.

Proof. Let $\xi = C\left(\frac{\log n}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right)^{\frac{1}{2}}$ and $C_3 = \frac{1}{\max\left(C_1, \frac{C_2}{3}\right)}$, using Bernstein's inequality we can write:

$$\mathbb{P}\left(|\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]| > \xi|\mathcal{M}\right) \leq 2\exp\left(-\frac{\frac{1}{2}\xi^2}{\frac{C_1}{\bar{N}|H|^{\frac{1}{2}}\lambda} + \frac{1}{3}\frac{C_2\xi}{\bar{N}|H|^{\frac{1}{2}}\lambda}}\right) \tag{26}$$

$$= 2\exp\left(-\frac{\frac{1}{2}C^2\log n}{C_1 + \frac{1}{3}C_2\xi}\right) \leq 2\exp\left(-\frac{C_3C^2\log n}{1+\xi}\right), \forall(\theta,t). \tag{27}$$

Therefore, if $C_4 > 0$ is given, and we choose $C^2 > \frac{3C_4}{C_3}$, then we can write:

$$\sup_{(\theta,t)\ \in\ \mathbb{R}^p\times\mathcal{I}} \mathbb{P}\left(|\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]| > C\left(\frac{\log n}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right)^{\frac{1}{2}}\right) \leq 2\exp\left(-\frac{3C_4\log n}{1+\xi}\right) \tag{28}$$

$$= 2n^{-\frac{3C_4}{1+\xi}}. \tag{29}$$

Now restricting to finite subsets $\mathcal{K}_n \subset \mathcal{K} \subset \mathbb{R}^p$ and $\mathcal{I}_n \subset \mathcal{I}$ where $\mathcal{K}_n \times \mathcal{I}_n$ has at most $\lfloor n^{\frac{2C_4}{1+\xi}} \rfloor$ elements, we have:

$$\mathbb{P}\left(\sup_{(\theta,t)\ \in\ \mathcal{K}_n\times\mathcal{I}_n} |\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]| > C\left(\frac{\log n}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right)^{\frac{1}{2}}\right) \leq 2n^{\frac{-C_4}{1+\xi}}. \tag{30}$$

From the H$\ddot{o}lder$-continuity of the estimator (since the two kernels have bounded first derivative):

$$\sup_{(\theta,t)\ \in\ \mathcal{K}\times\mathcal{I}} \{|\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]|\} - \sup_{(\theta,t)\ \in\ \mathcal{K}_n\times\mathcal{I}_n} \{|\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]|\} =$$

$$\left|\sup_{(\theta,t)\ \in\ \mathcal{K}\times\mathcal{I}} \{|\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]|\} - \sup_{(\theta,t)\ \in\ \mathcal{K}_n\times\mathcal{I}_n} \{|\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]|\}\right| \leq$$

$$\left|\sup_{(\theta,t)\ \in\ \mathcal{K}\times\mathcal{I}} \{\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]\} - \sup_{(\theta,t)\ \in\ \mathcal{K}_n\times\mathcal{I}_n} \{\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]\}\right| \leq$$

$$D||v^* - v_n^*||^{\alpha}, \tag{31}$$

where

$$v^* = \arg\sup_{(\theta,t)\ \in\ \mathcal{K}\times\mathcal{I}} \{\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]\}$$

$$v_n^* = \arg\sup_{(\theta,t)\ \in\ \mathcal{K}_n\times\mathcal{I}_n} \{\hat{p}(\theta,t) - \mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]\},$$

therefore:

$$\mathbb{P}\left(\sup_{(\theta,t)\,\in\,\mathcal{K}\times\mathcal{I}}|\hat{p}(\theta,t)-\mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]|>C\left(\frac{\log n}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right)^{\frac{1}{2}}+D||v^*-v_n^*||^{\alpha}\right)\leq$$
$$\mathbb{P}\left(\sup_{(\theta,t)\,\in\,\mathcal{K}_n\times\mathcal{I}_n}|\hat{p}(\theta,t)-\mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]|>C\left(\frac{\log n}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right)^{\frac{1}{2}}\right) \tag{32}$$

now, for sufficiently large C_4, $||v^*-v_n^*||\leq \frac{\sqrt{p+1}}{2}\sqrt[p+1]{\frac{(K_{max}-K_{min})^p(I_{max}-I_{min})}{\left\lfloor n^{\frac{2C_4}{1+\xi}}\right\rfloor}}$ and $D\left(\frac{\sqrt{p+1}}{2}\sqrt[p+1]{\frac{(K_{max}-K_{min})^p(I_{max}-I_{min})}{\left\lfloor n^{\frac{2C_4}{1+\xi}}\right\rfloor}}\right)^{\alpha}$ is negligible w.r.t. ξ as n tends to infinity, where K_{max} e K_{min} are the endpoints for each dimension of $\mathcal{K}$ (we assume them to be the same in each dimension for the sake of simplicity). Analogously for I_{max} and I_{min} (notice that $\mathcal{I}$ is monodimensional).

Therefore:

$$\mathbb{P}\Bigg(\sup_{(\theta,t)\,\in\,\mathcal{K}\times\mathcal{I}}|\hat{p}(\theta,t)-\mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]|>C\left(\frac{\log n}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right)^{\frac{1}{2}}+$$
$$D\left(\frac{p+1}{4}\right)^{\frac{\alpha}{2}}\left(\frac{(K_{max}-K_{min})^p(I_{max}-I_{min})}{\left\lfloor n^{\frac{2C_4}{1+\xi}}\right\rfloor}\right)^{\frac{\alpha}{p+1}}\Bigg)\leq$$
$$\mathbb{P}\left(\sup_{(\theta,t)\,\in\,\mathcal{K}_n\times\mathcal{I}_n}|\hat{p}(\theta,t)-\mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]|>C\left(\frac{\log n}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right)^{\frac{1}{2}}\right). \tag{33}$$

From (30) and (33), we can write:

$$\mathbb{P}\Bigg(\sup_{(\theta,t)\,\in\,\mathcal{K}\times\mathcal{I}}|\hat{p}(\theta,t)-\mathbb{E}[\hat{p}(\theta,t)|\mathcal{M}]|<C\left(\frac{\log n}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right)^{\frac{1}{2}}+$$
$$D\left(\frac{p+1}{4}\right)^{\frac{\alpha}{2}}\left(\frac{(K_{max}-K_{min})^p(I_{max}-I_{min})}{\left\lfloor n^{\frac{2C_4}{1+\xi}}\right\rfloor}\right)^{\frac{\alpha}{p+1}}\Bigg)\geq 1-2n^{\frac{-C_4}{1+\xi}} \tag{34}$$

Therefore, as $n\to\infty$ with probability 1:

$$|\hat{p}(\theta,t)-p(\theta,t)-O(\lambda)-O(tr(|H|))|=O\Bigg[C\left(\frac{\log n}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right)^{\frac{1}{2}}+$$
$$D\left(\frac{p+1}{4}\right)^{\frac{\alpha}{2}}\left(\frac{(K_{max}-K_{min})^p(I_{max}-I_{min})}{\left\lfloor n^{\frac{2C_4}{1+\xi}}\right\rfloor}\right)^{\frac{\alpha}{p+1}}\Bigg],\forall\,(\theta,t)\;\in\;\mathcal{K}\times\mathcal{I} \tag{35}$$

Finally, we get:

$$\hat{p}(\theta,t)=p(\theta,t)+O\left[\left(\frac{\log n}{\bar{N}|H|^{\frac{1}{2}}\lambda}\right)^{\frac{1}{2}}+\lambda+tr(H)\right],\forall\,(\theta,t)\;\in\;\mathcal{K}\times\mathcal{I} \tag{36}$$

□

B UPPER BOUND ON THE KL-DIVERGENCE BETWEEN THE PRIOR AND THE POSTERIOR

In this section, we report the steps needed to get an upper bound on the KL-Divergence between the posterior q our prior $\hat{p}$. Let us define $S = \frac{1}{a_0(-\rho)\bar{N}\lambda}\sum_{i=1}^{n}\sum_{j=1}^{M_i} K_T(\frac{t-t_i}{\lambda})$, hence:

$$D_{KL}(q||\hat{p}(\cdot,t)) = \int q(\theta) log \frac{q(\theta)}{\hat{p}(\theta,t)} d\theta = \int q(\theta) log \frac{q(\theta)}{\frac{S}{S}\hat{p}(\theta,t)} d\theta \tag{37}$$

$$= \int q(\theta) log \frac{q(\theta)}{\frac{1}{Sa_0(-\rho)\bar{N}|H|^{\frac{1}{2}}\lambda}\sum_{i=1}^{n} K_T(\frac{t-t_i}{\lambda})\sum_{j=1}^{M_i} K_S(H^{-\frac{1}{2}}(\theta-\theta_{ij}))} d\theta + \int q(\theta)\log\frac{1}{S} d\theta \tag{38}$$

Now the first term in Equation (38) is the KL-Divergence between two Mixture of Gaussians, which can be upper bounded using the same procedure as in Hershey and Olsen [2007], and the second term is a constant in the ELBO optimization. Therefore:

$$D_{KL}(q||\hat{p}(\cdot,t)) \leq D_{KL}(\chi^{(2)}||\chi^{(1)}) + \log\frac{1}{S} + \sum_{i,j}\chi^{(2)}_{j,i} D_{KL}(f^q_i||f^{\hat{p}}_j), \tag{39}$$

where we are rewriting $q = \sum_i c^q_i f^q_i$ and $\hat{p} = \sum_j c^{\hat{p}}_j f^{\hat{p}}_j$ with c^y_x being a generic weight and $f^y_x = \mathcal{N}(\mu^y_x, \Sigma^y_x)$ being a generic component, $(x,y) \in \{(i,q),(j,\hat{p})\}$. Furthermore, we have:

$$\chi^{(1)}_{i,j} = \frac{c^{\hat{p}}_j \chi^{(2)}_{j,i}}{\sum_{i'}\chi^{(2)}_{j,i'}}, \qquad \chi^{(2)}_{j,i} = \frac{c^{(q)}_i \chi^{(1)}_{i,j} e^{-D_{KL}(f^q_i||f^{\hat{p}}_j)}}{\sum_{j'}\chi^{(1)}_{i,j'} e^{-D_{KL}(f^q_i||f^{\hat{p}}_{j'})}}. \tag{40}$$

Finally, notice that $c^q_i = \frac{1}{C}$ for each i, where C is the number of components for the posterior, whereas $c^{\hat{p}}_j = \frac{1}{Sa_0(-\rho)\bar{N}\lambda} K_T(\frac{t-t_i}{\lambda})$, with a little abuse of notation over the index i and j.

C PROOF OF THEOREM 4.1

The prof of Theorem 4.1 is straightforward, we just need to follow the same procedure of Tirinzoni et al. [2018a] plugging in the bound on the KL-Divergence of Equation (39). In the following we report the proof for completeness.

Proof. We start from Lemma 2 of Tirinzoni et al. [2018a] with variational parameter $\hat{\xi} = (\hat{\mu}_1,\dots,\hat{\mu}_C,\hat{\Sigma}_1,\dots,\hat{\Sigma}_C)$, whereas, for the right-hand side, we set $\mu_i = \theta^*$ and $\Sigma_i = cI$ for each $i = 1,\dots,C$, for some $c > 0$:

$$\mathbb{E}_{q_{\hat{\xi}}}\left[\left|\left|\tilde{B}_\theta\right|\right|^2_\nu\right] \leq \inf_{\xi\in\Xi}\left\{\mathbb{E}_{q_\xi}\left[\left|\left|\tilde{B}_\theta\right|\right|^2_\nu\right] + \mathbb{E}_{q_{\hat{\xi}}}[\nu(\theta)] + 2\frac{\psi}{N} D_{KL}(q_\xi||\hat{p})\right\} + 8\frac{R^2_{max}}{(1-\gamma)^2}\sqrt{\frac{\log\frac{2}{\delta}}{2N}}$$

$$\leq \mathbb{E}_{\mathcal{N}(\theta^*,cI)}\left[\left|\left|\tilde{B}_\theta\right|\right|^2_\nu\right] + \mathbb{E}_{\mathcal{N}(\theta^*,cI)}[\nu(\theta)] + 2\frac{\psi}{N} D_{KL}(\mathcal{N}(\theta^*,cI)||\hat{p}) + 8\frac{R^2_{max}}{(1-\gamma)^2}\sqrt{\frac{\log\frac{2}{\delta}}{2N}}. \tag{41}$$

From Appendix B we have:

$$D_{KL}(\mathcal{N}(\theta^*,cI)||\hat{p}) \leq D_{KL}(\chi^{(2)}||\chi^{(1)}) + \log\frac{1}{S} + \sum_j \chi^{(2)}_j D_{KL}(\mathcal{N}(\theta^*,cI)||\mathcal{N}(\theta_j,\sigma^2 I)), \tag{42}$$

where

$$\chi^{(1)}_j = c^{\hat{p}}_j, \qquad \chi^{(2)}_j = \frac{c^{\hat{p}}_j e^{-D_{KL}(\mathcal{N}(\theta^*,cI)||\mathcal{N}(\theta_j,\sigma^2 I))}}{\sum_{j'} c^{\hat{p}}_{j'} e^{-D_{KL}(\mathcal{N}(\theta^*,cI)||\mathcal{N}(\theta_{j'},\sigma^2 I))}} \tag{43}$$

obtained noticing that we can remove the index i because we have reduced the posterior to one component. $\chi_j^{(2)}$ can be rewritten:

$$\chi_j^{(2)} = \frac{c_j^{\hat{p}} e^{-\frac{1}{2\sigma^2}||\theta^* - \theta_j||}}{\sum_{j'} c_{j'}^{\hat{p}} e^{-\frac{1}{2\sigma^2}||\theta^* - \theta_{j'}||}} \tag{44}$$

if we plug in the closed form expression of the KL-Divergence (45) into its definition.

$$D_{KL}(\mathcal{N}(\theta^*, cI)||\mathcal{N}(\theta_j, \sigma^2 I)) = \frac{1}{2}\left(p\log\frac{\sigma^2}{c} + p\frac{c}{\sigma^2} + \frac{||\theta^* - \theta_j||}{\sigma^2} - p\right). \tag{45}$$

Now we proceed upper bounding the first and then the third term of (42):

$$D_{KL}(\chi^{(2)}||\chi^{(1)}) = \sum_j \chi_j^{(2)} \log\frac{\chi_j^{(2)}}{\chi_j^{(1)}} \tag{46}$$

$$= \sum_j \chi_j^{(2)} \log\chi_j^{(2)} - \sum_j \chi_j^{(2)} \log\chi_j^{(1)} \tag{47}$$

$$\leq \sum_j \chi_j^{(2)} \log\frac{1}{c_j^{\hat{p}}} \tag{48}$$

where we got (48) just noticing in (47) that the first term is negative. Considering the third term, we have:

$$\begin{aligned}\sum_j \chi_j^{(2)} D_{KL}(\mathcal{N}(\theta^*, cI)||\mathcal{N}(\theta_j, \sigma^2 I)) &= \frac{1}{2}\sum_j \chi_j^{(2)}\left(p\log\frac{\sigma^2}{c} + p\frac{c}{\sigma^2} + \frac{||\theta^* - \theta_j||}{\sigma^2} - p\right) \\ &\leq \frac{1}{2}p\log\frac{\sigma^2}{c} + \frac{1}{2}p\frac{c}{\sigma^2} + \sum_j \chi_j^{(2)}\frac{||\theta^* - \theta_j||}{2\sigma^2}.\end{aligned} \tag{49}$$

Therefore:

$$\begin{aligned}&D_{KL}(\mathcal{N}(\theta^*, cI)||\hat{p}) \leq \\ &\qquad \sum_j \chi_j^{(2)} \log\frac{1}{c_j^{\hat{p}}} + \log\frac{1}{S} + \frac{1}{2}p\log\frac{\sigma^2}{c} + \frac{1}{2}p\frac{c}{\sigma^2} + \sum_j \chi_j^{(2)}\frac{||\theta^* - \theta_j||}{2\sigma^2}.\end{aligned} \tag{50}$$

Now leveraging the above equation, the following upper bound obtained in the proof of Theorem 3 in Tirinzoni et al. [2018a]:

$$\mathbb{E}_{\mathcal{N}(\theta^*, cI)}\left[\left|\left|\tilde{B}_\theta\right|\right|_\nu^2\right] \leq 2\left|\left|\tilde{B}_{\theta^*}\right|\right|_\nu^2 + \frac{1}{2}\gamma^2\kappa^2 c^2\phi_{max}^4 + c(\theta_{max}\phi_{max}(1+\gamma))^2, \tag{51}$$

and setting c=$\frac{1}{N}$ (since the bound hold for any constant parameter $c > 0$), $c_1 = \frac{8R_{max}^2}{\sqrt{2}(1-\gamma)^2}$, $c_2 = \theta_{max}^2\phi_{max}^2(1-\gamma)^2 + \psi p\log\sigma^2 + 2\psi\sum_j \chi_j^{(2)}\log\frac{1}{c_j^{\hat{p}}} + 2\psi\log\frac{1}{S}$, $c_3 = \frac{1}{2}\gamma^2\kappa^2\phi_{max}^4 + \frac{\psi p}{\sigma^2}$ and $\varphi(\Theta_s) = \frac{1}{\sigma^2}\sum_j \chi_j^{(2)}||\theta^* - \theta_j||$, we can rewrite Equation (41) in the following way:

$$\mathbb{E}_{q_\xi}\left[\left|\left|\tilde{B}_\theta\right|\right|_\nu^2\right] \leq 2\left|\left|\tilde{B}_{\theta^*}\right|\right|_\nu^2 + v(\theta^*) + c_1\sqrt{\frac{\log\frac{2}{\delta}}{N}} + \frac{c_2 + \psi p\log N + \psi\varphi(\Theta_s)}{N} + \frac{c_3}{N} \tag{52}$$

□

D EXPERIMENTAL DETAILS

In this section, we provide some additional experimental details together with further results.

D.1 PARAMETRIZATION

ADAM [Kingma and Ba, 2014] is used in every experiment as optimizer. The source tasks are solved by a direct minimization of the TD error as described in section 3.4 of Tirinzoni et al. [2018a], using a *batch size* of 50 for the rooms environments and of 32 for Mountain Car and the lake Como water system, a *buffer size* of 50000, the projection parameter of the mellow-max TD error gradient set to 0.5, the learning rate $\alpha = 10^{-3}$. The exploration is ϵ-greedy with ϵ linearly decaying from 1 to 0.01 for Mountain Car and to 0.02 for the rooms environments. Both decays happen within 50% of the maximum number of learning iterations. In the lake Como environment we used a soft-max (Gibbs) policy with parameter β linearly increasing from 0.5 to 9.275 through the learning iterations.

In the **rooms** environments, for what concern the two transfer algorithms, c-T2VT, and c-MGVT, we have the following parametrization: *batch size* of 50, *buffer size* of 50000, projection parameter of the mellow-max TD error gradient set to 0.5 (see section 3.4 of Tirinzoni et al. [2018a]), the parameter of Equation (2) $\psi = 10^{-6}$, 10 weights to estimate the expected TD error, the learning rates are set to $\alpha_\mu = 10^{-3}$ and $\alpha_L = 0.1$ for the mean and the Cholesky factor L of the posterior (moreover, the minimum eigenvalue reachable by L is set to $\sigma^2_{min} = 10^{-4}$). Finally, for the prior, we use a diagonal isotropic matrix $H = 10^{-5}I$ and $\lambda = 0.3333$ in the context of c-T2VT, furthermore, we have $\Sigma = 10^{-5}I$ for the prior in the context of c-MGVT.

In the **Mountain Car** environment, c-T2VT and c-MGVT are parametrized in the following way: *batch size* of 500, *buffer size* of 10000, projection parameter of the mellow-max TD error gradient set to 0.5, the parameter of Equation (2) $\psi = 10^{-4}$, 10 weights to estimate the expected TD error, the learning rates are set to $\alpha_\mu = 10^{-3}$ and $\alpha_L = 10^{-4}$ for the mean and the Cholesky factor L of the posterior (moreover, the minimum eigenvalue reachable by L is set to $\sigma^2_{min} = 10^{-4}$). Finally, for the prior, we use a diagonal isotropic matrix $H = 10^{-5}I$ and $\lambda = 0.3333$ in the context of c-T2VT, furthermore, we have $\Sigma = 10^{-5}I$ for the prior in the context of c-MGVT.

In the **lake Como** water system, 3-T2VT and 3-MGVT are parametrized in the following way: *batch size* of 32, *buffer size* of 10000, projection parameter of the mellow-max TD error gradient set to 0.5, the parameter of Equation (2) $\psi = 10^{-4}$, 4 weights to estimate the expected TD error, the learning rates are set to $\alpha_\mu = 10^{-3}$ and $\alpha_L = 10^{-4}$ for the mean and the Cholesky factor L of the posterior (moreover, the minimum eigenvalue reachable by L is set to $\sigma^2_{min} = 10^{-4}$). Finally, for the prior, we use a diagonal isotropic matrix $H = 10^{-5}I$ and λ was chosen through the maximum-likelihood approach of Section 6.5 in the context of 3-T2VT, furthermore, we have $\Sigma = 10^{-5}I$ for the prior in the context of 3-MGVT.

D.2 TEMPORAL DYNAMICS

In this section, we provide the analytical form of the different dynamics employed in our experiments. Notice that these dynamics need to be plugged into the mean of our Gaussian distribution from where we sample the parametrization defining the task (for the rooms environment we will sample the positions of the doors, whereas, for the Mountain Car environment, we will sample the base speed).

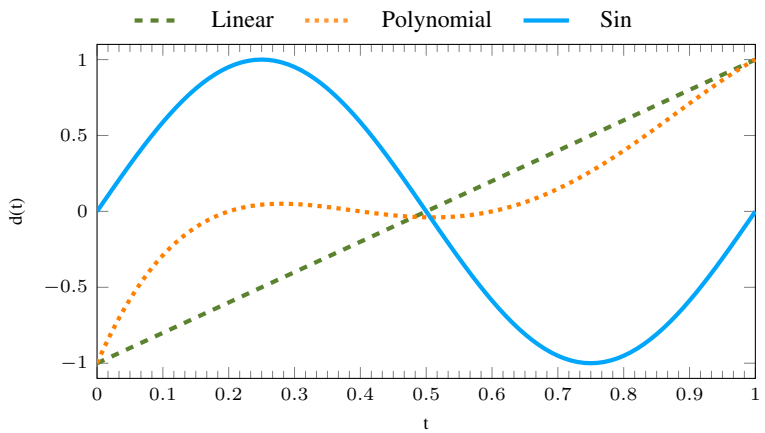


Figure 1: Temporal dynamics.

- **Linear**: $2t - 1,\ t \in [0, 1]$;
- **Polynomial**: $at^4 + bt^3 + ct^2 + dt + e,\ t \in [0, 1]$ and $a = -15.625$, $b = 39.5833$, $c = -31.875$, $d = 9.91667$ and $e = -1$;
- **Sinusoidal**: $\sin(2\pi t),\ t \in [0, 1]$.

In Figure 1, we report the graphical representation of the above analytical functions.

Now, given the range for a parameter $[k_{min}, k_{max}]$, a given dynamic will span over this interval in the following way: $d(t)\frac{(k_{max}-k_{min})}{2} + \frac{(k_{max}+k_{min})}{2}$. Finally, notice that, $[k_{min}, k_{max}] = [0.001, 0.0015]$ for Mountain Car, whereas $[k_{min}, k_{max}] = [0.7 + padding, 9.3 - padding]$ for the parameters of the rooms environments. The $padding$ variable is 0 for the 2-rooms, whereas is 2 for the 3-rooms environments. This $padding$ variable was necessary in the 3-rooms environments in order for the TD gradient algorithm to be able to solve the source tasks in every configuration of the two doors.

D.3 λ-SENSITIVITY RESULTS

In Figures 2 and 3, we report a sensitivity analysis of our algorithm w.r.t. λ in the 2-rooms environment. This analysis is carried out computing the performance of the learning algorithm w.r.t. different values of the previously mentioned parameter (whereas $H = 10^{-5}I$ for every λ). These results are also compared with the performance of the algorithm when λ is chosen according to the likelihood optimization described in Section 6.5. In Figures 4 and 5, we report the above-described analysis in the context of the 3-rooms environment, whereas, in Figures 6 and 7, we have the Mountain Car environment.

In the context of both rooms environments, the performance of the likelihood approach is satisfying, for both 1-T2VT and 3-T2VT, even though in some cases it is not optimal. For what concern, the polynomial dynamic this may be due to its plateau (see Figure 1) which bias the choice for λ toward bigger values since the likelihood is evaluated in a cross-validation manner. For the same reason, in the $\sin$ dynamic case, the likelihood-based approach tends to select an average λ. Finally, the linear case in the 2-rooms is almost optimal, whereas, in the 3-rooms, the performance decreases. This is due to the fact that, in the 3-rooms environment, we have 2 parameters governing the dynamics (the two doors positions) making the choice of λ harder to make in this setting.

In the context of the Mountain Car environment the likelihood approach always choose the best λ as shown in Figures 6 and 7.

Implementation Details: since the $\lambda \in [0, 1]$, we performed a grid search in order to optimize Equation (5) .

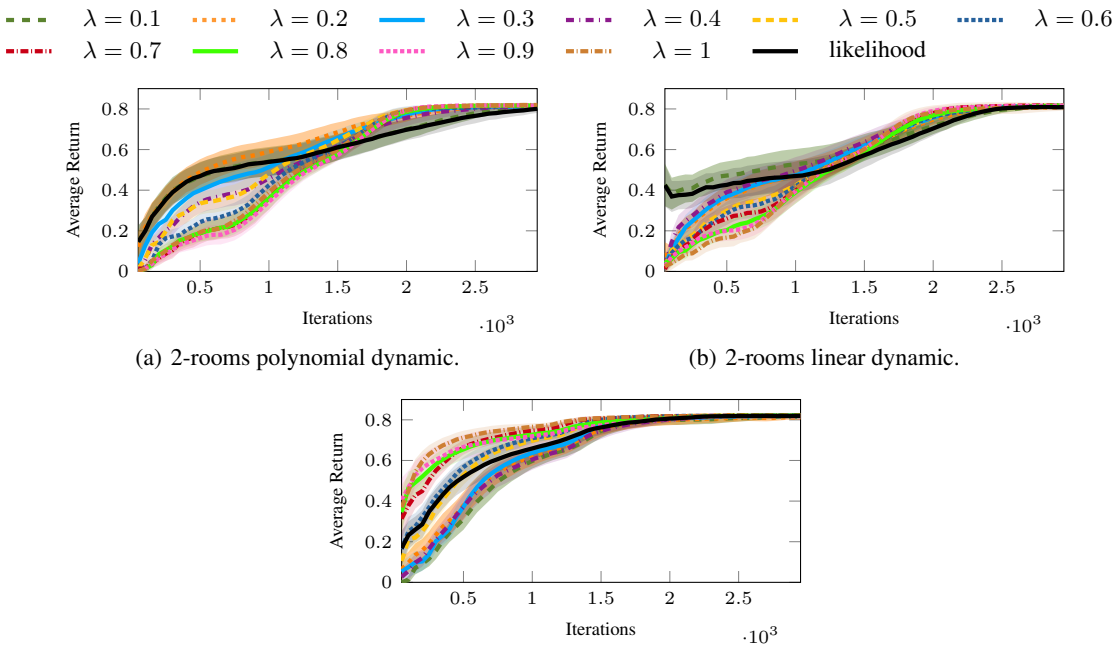
D.4 FURTHER ENVIRONMENTAL SETTINGS: MOUNTAIN CAR AND LAKE COMO WATER SYSTEM

D.4.1 Mountain Car

The state space consists in the *position* and *velocity* of the car. The reward function is always -1 so the agent must reach the goal as soon as possible. The available actions are *backward full throttle*, *zero throttle* and *forward full throttle* encoded as $[-1, 0, 1]$. The discount factor is $\gamma = 0.99$. The goal position is 0.5. Finally, the transition function is $position_{t+1} = position_t + velocity_{t+1}$, $velocity_{t+1} = velocity_t + a_t * 0.001 - 0.0025 * \cos(3 * position_t)$. The *velocity* is clipped whenever exits the range $[-0.07, 0.07]$ the *position* is bound to lie in $[-1.2, 0.6]$.

D.4.2 Lake Como

The reward function in the lake Como water system is composed of three main costs. The demand cost is a squared function of the discrepancy between actual release and water demand: $-4(\varrho_{t+1} - demand_t)^2$ if t is between may and august, otherwise $-(\varrho_{t+1} - demand_t)^2$. The flooding cost is a constant penalty inflicted to the agent whenever a water level flooding threshold is broken: $-1\ if\ water\ level > 1.24\ else\ 0$. Finally, the unfeasibility penalty is just a discrepancy between the action requested by the agent and the actual release the system was able to accomplish: $-|a_t - \varrho_{t+1}|$. Each component is rescaled in $[-1, 0]$ and contribute uniformly for $\frac{1}{3}$ to the reward function. The actions available to the agent are 8 different amount of water to be released: $[0, 79.39, 88.10, 110.39, 148.39, 200.13, 225.25, 491.61]$. The discount factor is $\gamma = 0.9999$.

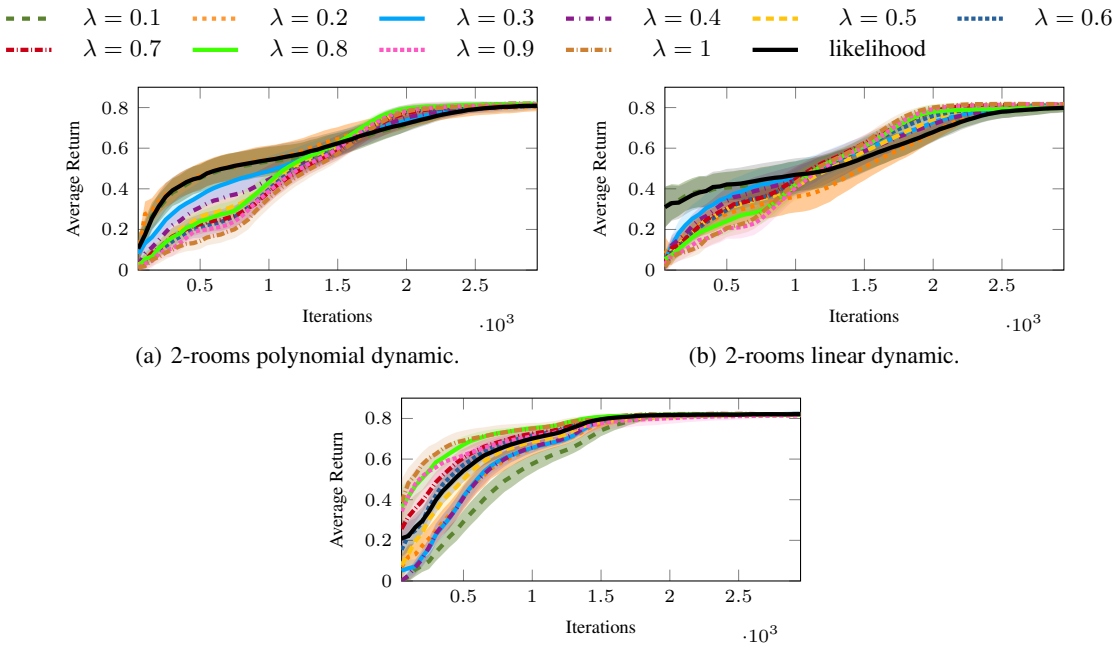


(a) 2-rooms polynomial dynamic.

(b) 2-rooms linear dynamic.

(c) 2-rooms sin dynamic.

Figure 2: Average return achieved by 1-T2VT w.r.t. different choices of λ with 95% confidence intervals computed using 50 independent runs.

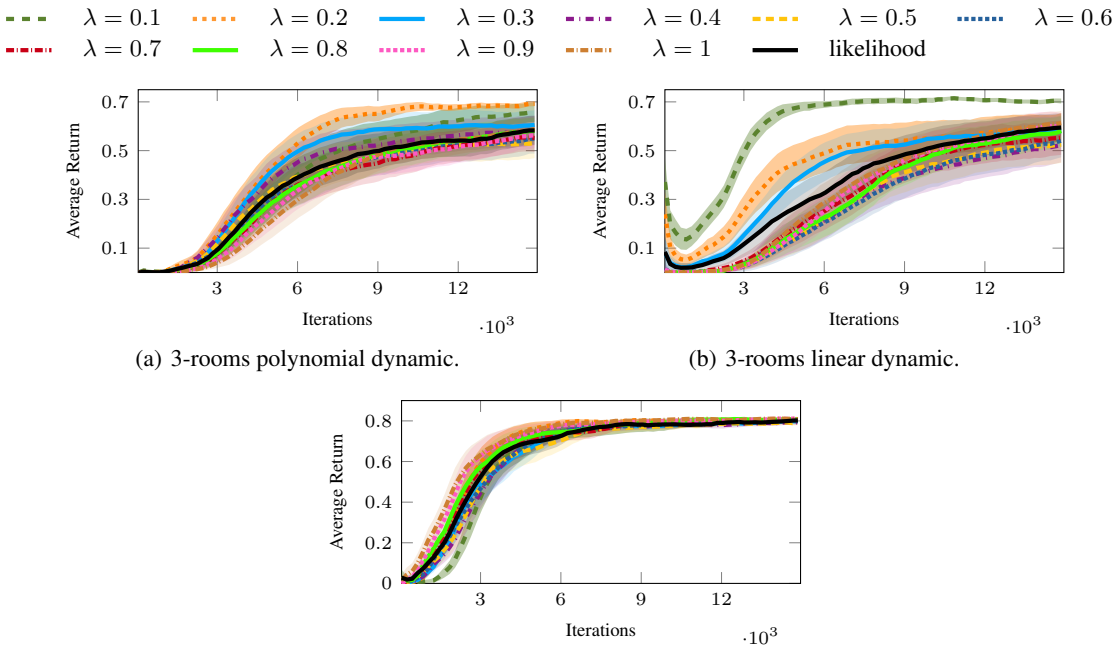


(a) 2-rooms polynomial dynamic.

(b) 2-rooms linear dynamic.

(c) 2-rooms sin dynamic.

Figure 3: Average return achieved by 3-T2VT w.r.t. different choices of λ with 95% confidence intervals computed using 50 independent runs.

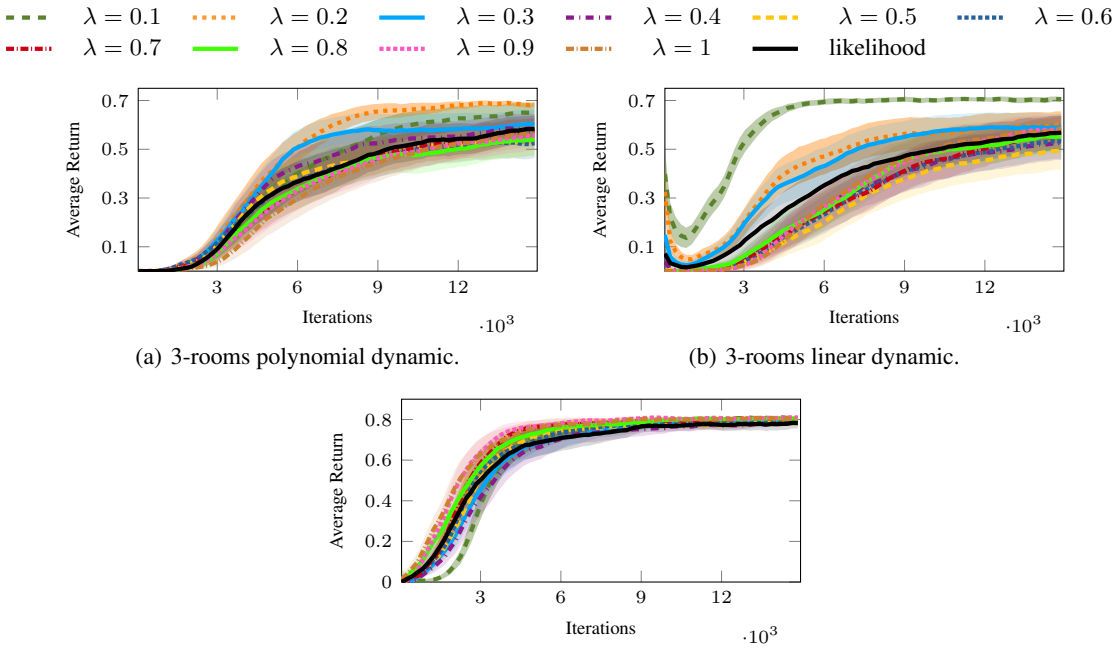


(a) 3-rooms polynomial dynamic.

(b) 3-rooms linear dynamic.

(c) 3-rooms sin dynamic.

Figure 4: Average return achieved by 1-T2VT w.r.t. different choices of λ with 95% confidence intervals computed using 50 independent runs.

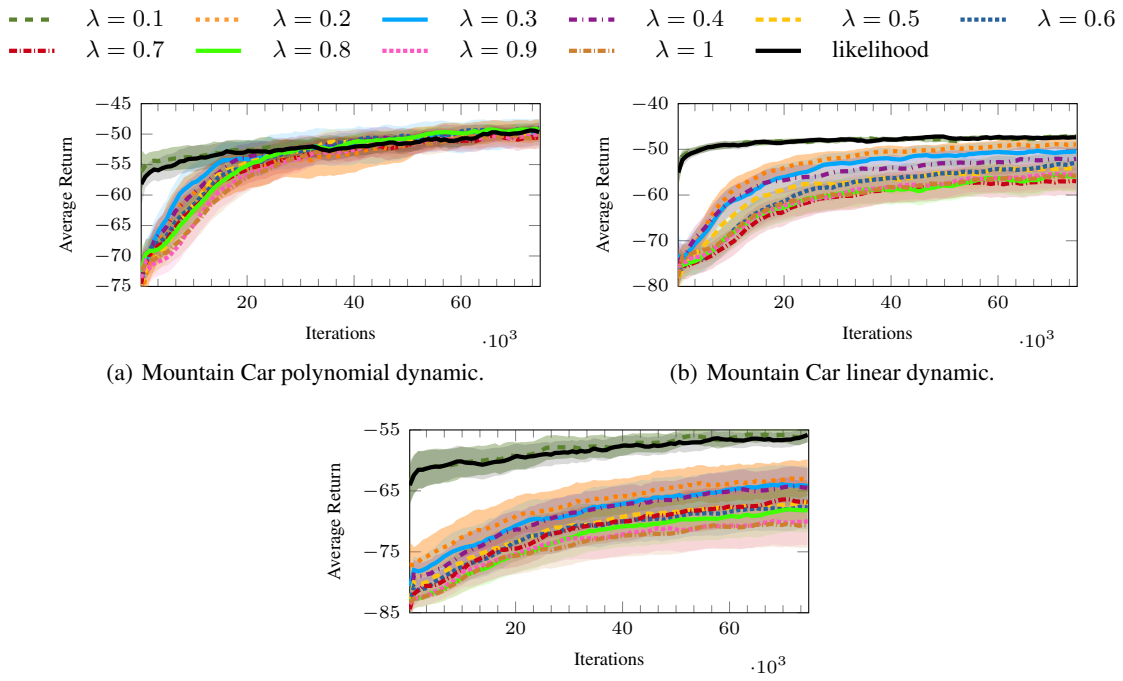


(a) 3-rooms polynomial dynamic.

(b) 3-rooms linear dynamic.

(c) 3-rooms sin dynamic.

Figure 5: Average return achieved by 3-T2VT w.r.t. different choices of λ with 95% confidence intervals computed using 50 independent runs.

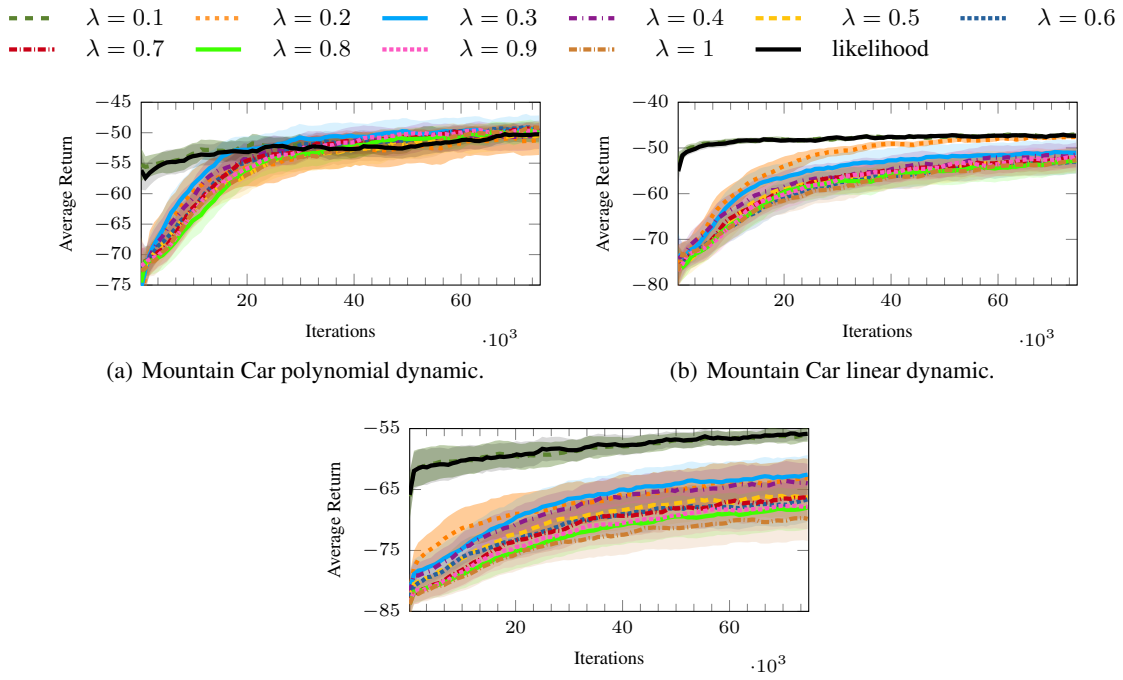


(a) Mountain Car polynomial dynamic.

(b) Mountain Car linear dynamic.

(c) Mountain Car sin dynamic.

Figure 6: Average return achieved by 1-T2VT w.r.t. different choices of λ with 95% confidence intervals computed using 50 independent runs.



(a) Mountain Car polynomial dynamic.

(b) Mountain Car linear dynamic.

(c) Mountain Car sin dynamic.

Figure 7: Average return achieved by 3-T2VT w.r.t. different choices of λ with 95% confidence intervals computed using 50 independent runs.