THE EUROPEAN CONFERENCE ON MACHINE LEARNING & PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES

ECML PKDD 2017

Bayesian Inference for Least-Squares Temporal Difference Regularization

Nikolaos Tziortziotis[†] and Christos Dimitrakakis[⋆]

[†] DaSciM, LIX, École Polytechnique, France

[⋆] University of Lille, France - SEAS, Harvard University, USA

✉ email: {ntziorzi, christos.dimitrakakis}@gmail.com

Abstract

Bayesian Least-Squares Temporal Difference:

✓ Fully Bayesian approach for LSTD

✓ Probabilistic inference of value functions

✓ Quantifies our uncertainty about value function

Variational Bayesian LSTD:

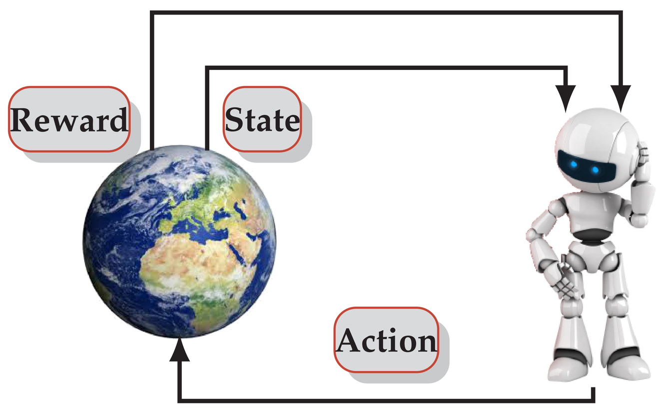
✓ Sparse model - Good generalisation capabilities

✓ Automatically determine the model's complexity

✓ No need to select a regularization parameter

Reinforcement Learning (RL)

Learning to act in an unknown environment, by interaction and reinforcement.



RL tasks folmulated as **MDPs**, $\{\mathcal{S}, \mathcal{A}, P, r, \gamma\}$

Policy π: $\mathcal{S} \to \mathcal{A}$ (map states to actions)

Value function: $V^\pi(s) \triangleq \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t) | s_0 = s\right]$

Bellman operator:

$$(T^\pi V)(s) = r(s) + \gamma \int_{\mathcal{S}} V(s') dP(s'|s, \pi(s))$$

V is the unique fixed-point of Bellman operator:

$$V^\pi = T^\pi V^\pi \Rightarrow V^\pi = (\boldsymbol{I} - \gamma P^\pi)^{-1} \boldsymbol{r}$$

☹ The value function cannot be represented in an explicit way (continuous state space)

☹ The model of the MDP is unknown

☺ **Value function approximation:**

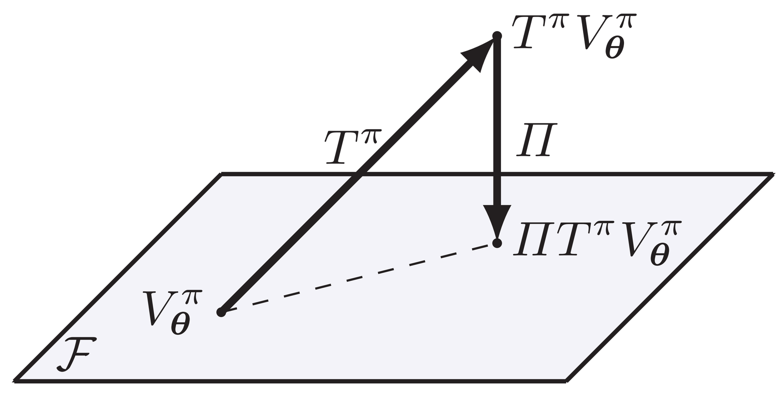
$$V_{\boldsymbol{\theta}}^\pi(s) = \boldsymbol{\phi}(s)^\top \boldsymbol{\theta} = \sum_{i=1}^{k} \phi_i(s)\theta_i,$$

where $\mathcal{F} = \{f_{\boldsymbol{\theta}} | f_{\boldsymbol{\theta}}(\cdot) = \boldsymbol{\phi}(\cdot)^\top \boldsymbol{\theta}\}$.

☺ **Access to a set of transitions:** $\mathcal{D} = \{(s_i, r_i, s_i')\}_{i=1}^n$, where we define: $\tilde{\Phi} = [\phi(s_1)^\top; \ldots; \phi(s_n)^\top]$, $\tilde{R} = [r_i, \ldots, r_n]^\top$ and $\tilde{\Phi}' = [\phi(s_1')^\top; \ldots; \phi(s_n')^\top]$.

Least-Squares Temporal Difference

Minimize the *mean-square projected Bellman error* (MSPBE): $\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^k} \|V_{\boldsymbol{\theta}}^\pi - \Pi T^\pi V_{\boldsymbol{\theta}}^\pi\|_D^2$.



Projection operator over $\mathcal{F}$: $\Pi = \Phi C^{-1} \Phi^\top D$

Nested Optimization Problem

$$\boldsymbol{u}^* = \arg\min_{\boldsymbol{u} \in \mathbb{R}^k} \|\Phi \boldsymbol{u} - T^\pi \Phi \boldsymbol{\theta}\|_D^2 \quad \textit{(Projection step)}$$

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^k} \|\Phi \boldsymbol{\theta} - \Phi \boldsymbol{u}^*\|_D^2 \quad \textit{(Fixed-point step)}$$

Solution

$$\boldsymbol{u}^* = \tilde{C}^{-1} \tilde{\Phi}(\tilde{R} + \gamma \tilde{\Phi}' \boldsymbol{\theta}),$$

$$\boldsymbol{\theta} = (\tilde{\Phi}^\top (\tilde{\Phi} - \gamma \tilde{\Phi}'))^{-1} \tilde{\Phi}^\top \tilde{R} = A^{-1} \boldsymbol{b},$$

where $\tilde{C} \triangleq \tilde{\Phi}^\top \tilde{\Phi}$, $A \triangleq \tilde{\Phi}^\top (\tilde{\Phi} - \gamma \tilde{\Phi}')$, and $\boldsymbol{b} \triangleq \tilde{\Phi}^\top \tilde{R}$.

As the number of samples n increases, the LSTD solution $\tilde{\Phi}\boldsymbol{\theta}$ converges to the fixed-point of $\hat{\Pi} T^\pi$.

Regularized LSTD Schemes

ℓ_2-LSTD	LARS-TD	$\ell_{2,1}$-LSTD
Lasso-TD	ℓ_1-PBR	Dantzig-LSTD
LC-TD	$\ell_{2,2}$-LSTD	ODDS-TD

Bayesian LSTD

Empirical Bellman operator:

$$\hat{T}^\pi V_{\boldsymbol{\theta}}^\pi = \boldsymbol{r} + \gamma P^\pi V_{\boldsymbol{\theta}}^\pi + N, \quad N \sim \mathcal{N}(0, \beta^{-1} \boldsymbol{I})$$

Given observations $\mathcal{D}$ (LSTD solution):

$$V_{\boldsymbol{\theta}}^\pi = \hat{\Pi} \hat{T}^\pi V_{\boldsymbol{\theta}}^\pi \Leftrightarrow \tilde{\Phi}^\top \tilde{R} = \tilde{\Phi}^\top (\tilde{\Phi} - \gamma \tilde{\Phi}') \boldsymbol{\theta} + \tilde{\Phi}^\top N$$

Linear regression model: $\boldsymbol{b} = A\boldsymbol{\theta} + \tilde{\Phi}^\top N$

Likelihood function: $p(\boldsymbol{b}|\boldsymbol{\theta}, \beta) = \mathcal{N}(\boldsymbol{b}|A\boldsymbol{\theta}, \beta^{-1}\tilde{C})$

✓ Maximum likelihood inference corresponds to standard LSTD solution

Prior distribution over $\boldsymbol{\theta}$: $p(\boldsymbol{\theta}|\alpha) = \mathcal{N}(\boldsymbol{\theta}|0, \alpha^{-1}\boldsymbol{I})$.

Logarithm of posterior distribution:

$$\ln p(\boldsymbol{\theta}|\mathcal{D}) \propto -\frac{\beta}{2} E_{\mathcal{D}}(\boldsymbol{\theta}) - \frac{\alpha}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta},$$

where $E_{\mathcal{D}}(\boldsymbol{\theta}) = (\boldsymbol{b} - A\boldsymbol{\theta})^\top \tilde{C}^{-1} (\boldsymbol{b} - A\boldsymbol{\theta})$ (MSPBE).

✓ Maximum a posteriori inference corresponds to ℓ_2 regularization ($\lambda = \alpha/\beta$)

Posterior distribution: $p(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}, S)$

$$S = (\alpha \boldsymbol{I} + \beta \underbrace{A^\top \tilde{C}^{-1} A}_{\Sigma})^{-1} \text{ and } \boldsymbol{m} = \beta S A^\top \tilde{C}^{-1} \boldsymbol{b}$$

Predictive distribution:

$$p(V_{\boldsymbol{\theta}}^\pi(s^*)|s^*, \mathcal{D}) = \int_{\boldsymbol{\theta}} p(V_{\boldsymbol{\theta}}^\pi(s^*)|\boldsymbol{\theta}, s^*) dp(\boldsymbol{\theta}|\boldsymbol{b}, \alpha, \beta)$$

$$= \mathcal{N}(V_{\boldsymbol{\theta}}^\pi(s^*)|\boldsymbol{\phi}(s^*)^\top \boldsymbol{m}, \boldsymbol{\phi}(s^*)^\top S \boldsymbol{\phi}(s^*))$$

Variational Bayesian LSTD

Prior distribution: $p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \prod_{i=1}^{k} \mathcal{N}(\theta_i|0, \alpha_i^{-1})$

Hyperprior over $\boldsymbol{\alpha}$: $p(\boldsymbol{\alpha}) = \prod_{i=1}^{k} \mathcal{Gamma}(\alpha_i|h_a, h_b)$

Hyperprior over β: $p(\beta) = \mathcal{Gamma}(\beta|h_c, h_d)$

Posterior distribution over $\mathcal{Z} = \{\boldsymbol{\theta}, \boldsymbol{\alpha}, \beta\}$:

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \beta|\boldsymbol{b}) = \frac{p(\boldsymbol{b}|\boldsymbol{\theta}, \beta)p(\beta)p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})}{p(\boldsymbol{b})}$$

☹ **Marginal likelihood is analytically intractable**

✓ **We resolt to variational inference**

Variational distribution: $\mathcal{Q}(\mathcal{Z}) = \mathcal{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta})\mathcal{Q}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})\mathcal{Q}_{\beta}(\beta)$

Optimal distribution for each factor:

$$\mathcal{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}, S)$$

$$\mathcal{Q}_{\beta}(\beta) = \mathcal{Gamma}(\beta|\tilde{c}, \tilde{d})$$

$$\mathcal{Q}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \prod_{i=1}^{k} \mathcal{Gamma}(\alpha_i|\tilde{a}_i, \tilde{b}_i)$$

where,

$$S = (diag\, \mathbb{E}[\boldsymbol{\alpha}] + \mathbb{E}[\beta]\Sigma)^{-1}, \quad \boldsymbol{m} = \mathbb{E}[\beta] S A^\top \tilde{C}^{-1} \boldsymbol{b},$$

$$\tilde{a}_i = h_a + \frac{1}{2}, \quad \tilde{b}_i = h_b + \frac{1}{2}\mathbb{E}[\theta_i^2],$$

$$\tilde{c} = h_c + \frac{k}{2}, \quad \tilde{d} = h_d + \frac{1}{2}\|\boldsymbol{b} - A\boldsymbol{m}\|_{\tilde{C}}^2 + \frac{1}{2}\mathrm{tr}(\Sigma S).$$

Lower bound

$$\mathcal{L}(\mathcal{Q}) = \frac{1}{2}\ln|S| - \frac{1}{2}|\tilde{C}| + \sum_{i=1}^{k}\{\ln\Gamma(\tilde{a}_i) - \tilde{a}_i \ln \tilde{b}_i\}$$

$$+ \ln\Gamma(\tilde{c}) - \tilde{c}\ln\tilde{d} + \frac{k}{2}(1 - \ln 2\pi) - k\ln\Gamma(h_a)$$

$$+ kh_a \ln h_b - \ln\Gamma(h_c) + h_c \ln h_d$$

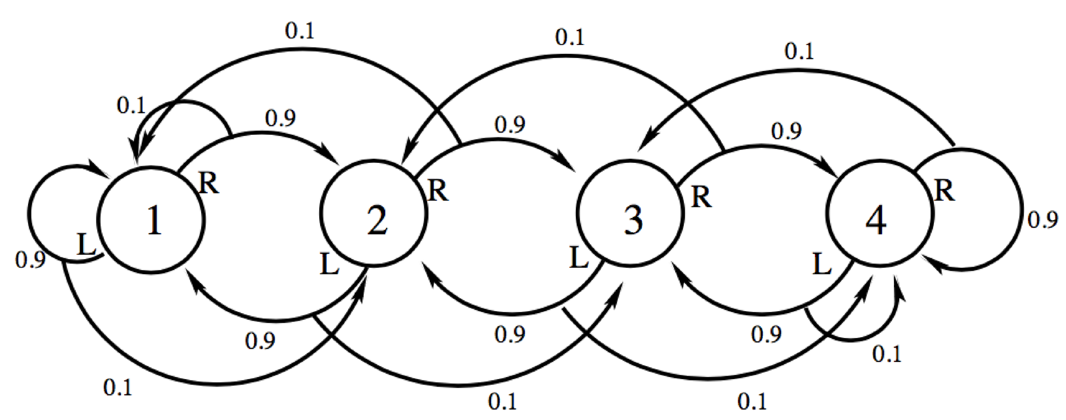
Experimental Results

Corrupted Chain

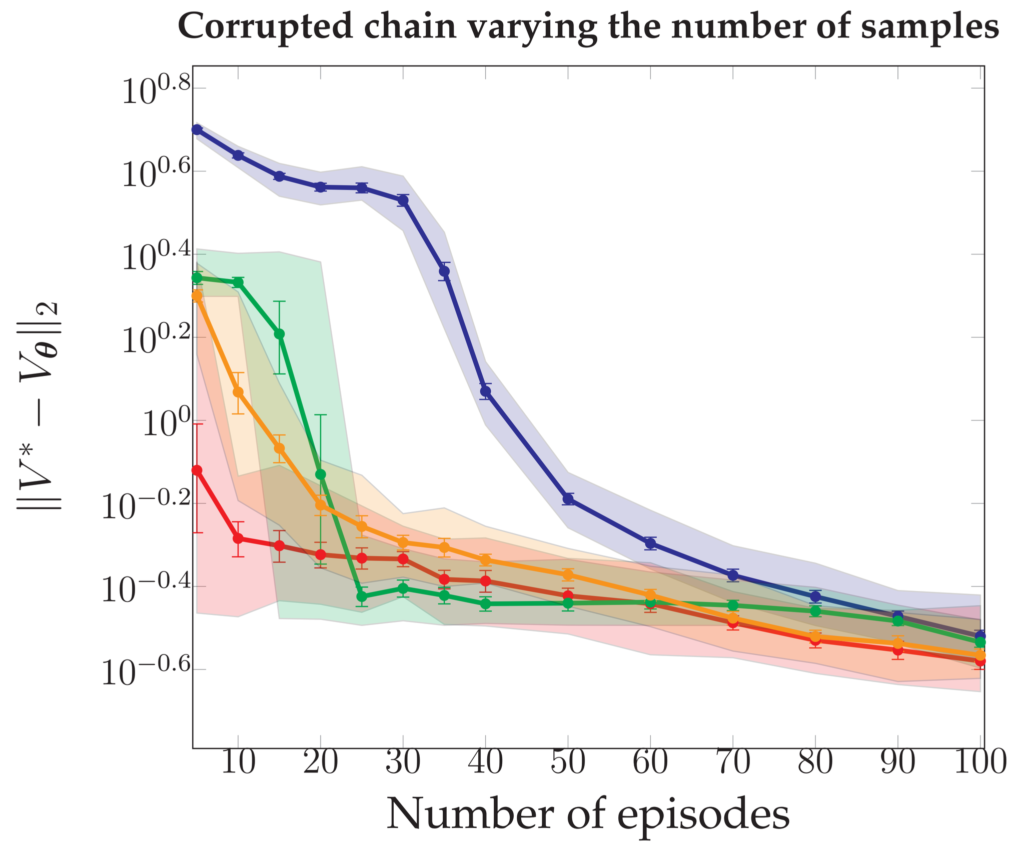
- A 20-state, 2-actions (*left* or *right*) MDP
- The probability of action's success is equal to 0.9
- A reward of 1 is given only at the ends of chain
- The horizon of each episode is set equal to 20
- Optimal policy: i) $s <= 10$: a = L, ii) $s > 10$: a = R
- Value fuction representation:

$$\phi(s) = (1, \mathrm{RBF}_1(s), \ldots, \mathrm{RBF}_5(s), x_1, \ldots, x_{600})$$

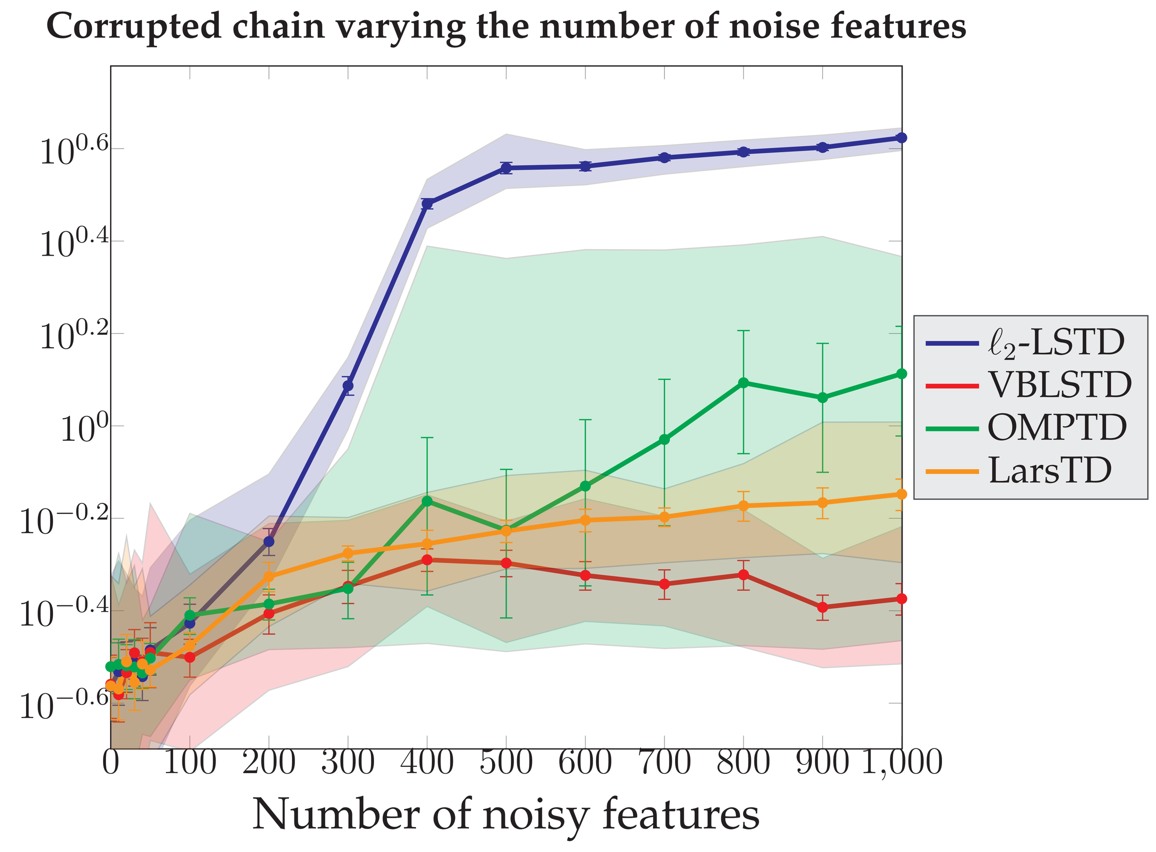
where $x_i \sim \mathcal{N}(0, 1), \forall i \in \{1, \ldots, 600\}$

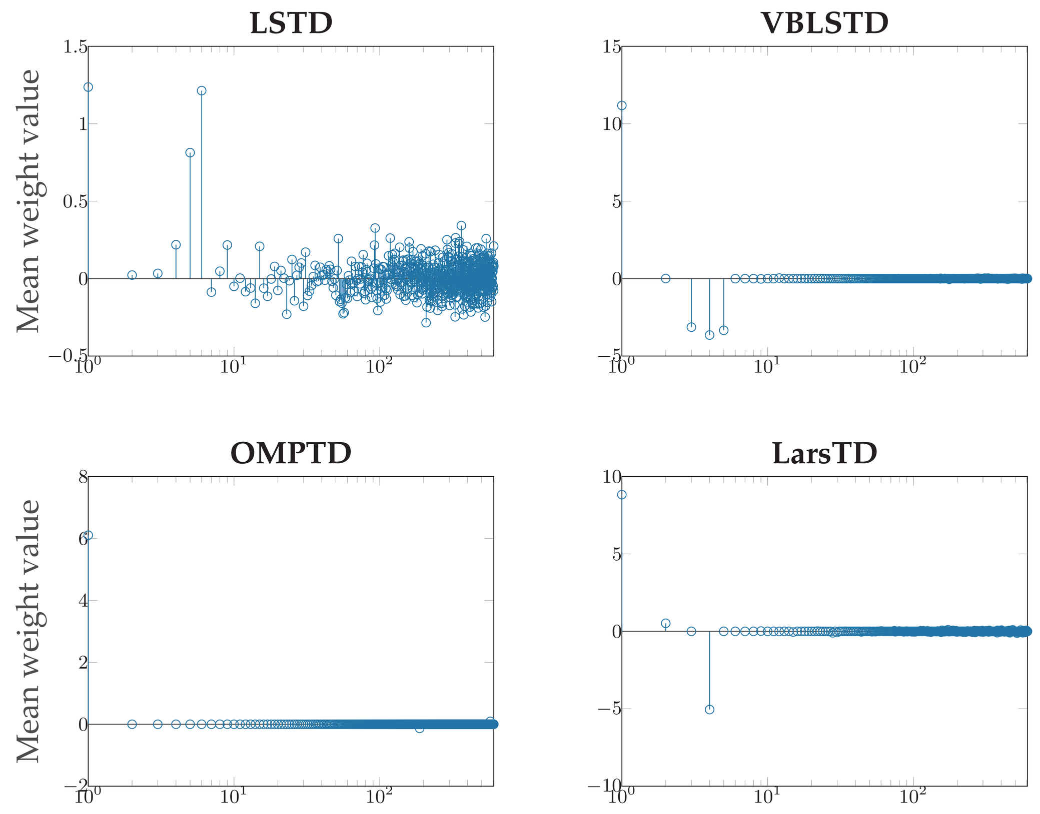


Corrupted chain varying the number of samples

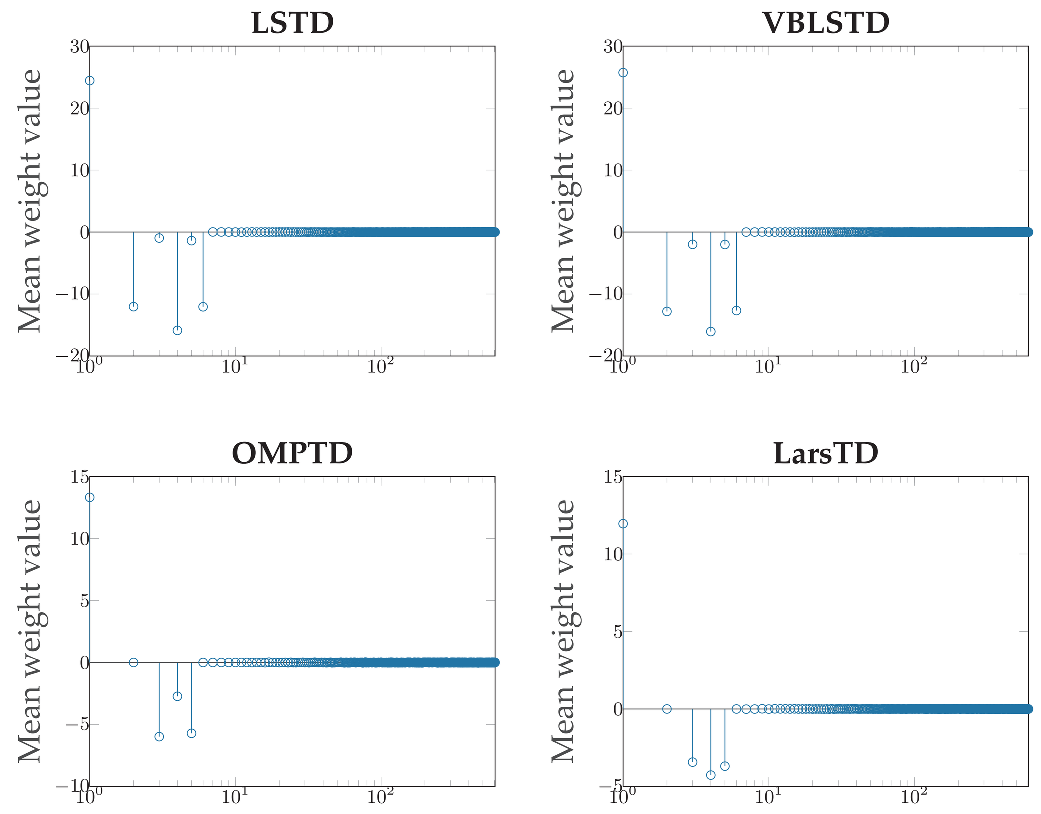


Corrupted chain varying the number of noise features





10 Training episodes



100 Training episodes

The 606 mean weight values. The first weight is the bias term, the next 5 correspond to the relevant features (RBFs), and the rest 600 correspond to the noise (irrelevant) features.