LSTD on Sparse Spaces

Yuri Grinberg, Mahdi Milani Fard, Joelle Pineau
School of Computer Science
McGill University
Montreal, Canada
{ygrinb,mmilan1,jpineau}@cs.mcgill.ca

1 Introduction

Efficient model selection and value function approximation are tricky tasks in reinforcement learning (RL), when dealing with large feature spaces. Even in batch settings, when the number of observed trajectories is small and the feature set is high-dimensional, there is little hope that we can learn a good value function directly based on all the features. To get better convergence and handle the overfitting problem, one has to make assumptions to reduce the complexity of the hypothesis space. This is a typical case of bias–variance tradeoff, where we bias our estimate to simpler models with lower complexity. The amount of this bias should be chosen according to the size of the data available to fit the model.

When very little is known about the RL environment, one can only make general-purpose simplifying assumptions that would hold in most environments. These include sparsity or smoothness of the value function in the feature space. Smoothness assumptions are often imposed on the learning task through direct regularization of the function approximator (e.g. with L^2 regularization [1, 2]). The sparsity assumption used in RL, on the other hand, assumes that most features are irrelevant in the value function and thus should be pruned either implicitly (e.g. with L^1 regularization [3]), or explicitly through feature generation methods from a large set of candidates (e.g. [4, 5, 6]).

In this paper, we consider another type of sparsity assumption that is commonly used in the compressed sensing (CS) literature. The original goal of CS was to compress a signal to lower dimensions (sample input at lower rates), such that the original can be almost perfectly reconstructed from the compressed version [7, 8]. This compression is usually done via a carefully chosen linear projection from the original space to a lower dimensional space. The sparsity assumption here, is that the signal itself is sparse in some known or unknown basis. This is fundamentally different from the typical sparsity assumption explained above (it is neither stronger, nor weaker). The types of methods used with this assumptions and the types of theoretical guarantees, will thus be somewhat different.

Random projections (borrowed from works on compressed sensing) have been studied in the context of feature extraction in RL for value function approximation. Ghavamzadeh et al. [9] have recently applied random projections to derive RL algorithms for feature selection in high-dimensional state spaces. They provide on-sample error bounds for the least-squares temporal difference (LSTD) algorithm, when the method is applied on the compressed space induced by the random projection. The method is shown to reduces the estimation error at the price of a controlled approximation error. Their work, however, does not use any sparsity assumption on the original feature space.

In this paper, we study the effect of random projections along with the sparsity assumptions reminiscent of those in the CS literature. We include a bias–variance analysis of LSTD when random projections are applied on sparse input signals. We show that the sparsity assumption let us extend on-sample error bounds of [9] to worst case bound on the entire state space.

2 Notations and Sparsity Assumption

Throughout this paper, column vectors are represented by lower case bold letters, and matrices are represented by bold capital letters. $|.|$ denotes the size of a set, and $\|.\|_0$ is Donoho's zero "norm" indicating the number of non-zero elements in a vector. $\|.\|$ denotes the L^2 norm for vectors and the operator norm for matrices: $\|\mathbf{M}\| = \sup_{\mathbf{v}} \|\mathbf{M}\mathbf{v}\|/\|\mathbf{v}\|$. Also, we denote the Moore-Penrose pseudo-inverse of a matrix $\mathbf{M}$ with $\mathbf{M}^\dagger$ and the smallest singular value of $\mathbf{M}$ by $\sigma_{\min}^{(M)}$.

We will be working in sparse feature spaces. Our state is represented by a vector $\mathbf{x} \in \mathcal{X}$ of D features, having $\|\mathbf{x}\| \leq 1$. We assume that $\mathbf{x}$ is k-sparse in some known or unknown basis $\mathbf{\Psi}$, implying that $\mathcal{X} \triangleq \{\mathbf{\Psi}\mathbf{z}, \text{ s.t. } \|\mathbf{z}\|_0 \leq k \text{ and } \|\mathbf{z}\| \leq 1\}$.

3 Random Projections and Inner Product

It is well known that random projections of appropriate sizes preserve enough information for exact reconstruction with high probability (see e.g. [10, 11]). In this section, we show that a function (almost-)linear in the original space is almost linear in the projected space, when we have random projections of appropriate sizes.

There are several types of random projection matrices that can be used. In this work, we assume that each entry in a projection $\mathbf{\Phi}^{D\times d}$ is an i.i.d. sample from a Gaussian [1]:

$$\phi_{i,j} = \mathcal{N}(0, 1/d). \tag{1}$$

We build our work on the following (based on theorem 4.1 from [10]), which shows that for a finite set of points, inner product with a fixed vector is almost preserved after a random projection.

Theorem 1. *Let $\mathbf{\Phi}^{D\times d}$ be a random projection according to Eqn 1. Let S be a finite set of points in $\mathbb{R}^D$. Then for any fixed $\mathbf{w}$ and $\epsilon > 0$:*

$$\forall \mathbf{s} \in S : \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{s}\rangle - \langle \mathbf{w}, \mathbf{s}\rangle\right| \leq \epsilon\|\mathbf{w}\|\|\mathbf{s}\|, \tag{2}$$

fails with probability less than $(4|S| + 2)e^{-d\epsilon^2/48}$.

The above theorem is based on the well-known Johnson–Lindenstrauss lemma (see [10]), which considers random projections of finite sets of points. We derive the corresponding theorem for sparse feature spaces.

Theorem 2. *Let $\mathbf{\Phi}^{D\times d}$ be a random projection according to Eqn 1. Let $\mathcal{X}$ be a D-dimensional k-sparse space. Then for any fixed $\mathbf{w}$ and $\epsilon > 0$:*

$$\forall \mathbf{x} \in \mathcal{X} : \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{x}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle\right| \leq \epsilon\|\mathbf{w}\|\|\mathbf{x}\|, \tag{3}$$

fails with probability less than:

$$(eD/k)^k(4(12/\epsilon)^k + 2)e^{-d\epsilon^2/192} \leq e^{k\log(12eD/\epsilon k) - d\epsilon^2/192 + \log 5}.$$

Note that the above theorem does not require $\mathbf{w}$ to be in the sparse space, and thus is different from guarantees on the preservation of inner product between vectors in the sparse space.

Proof of Theorem 2. The proof follows the steps of the proof of theorem 5.2 from [12]. Because $\mathbf{\Phi}$ is a linear transformation, we only need to prove the theorem when $\|\mathbf{w}\| = \|\mathbf{x}\| = 1$.

Denote $\mathbf{\Psi}$ to be the basis with respect to which $\mathcal{X}$ is sparse. Let $T \subset \{1, 2, \ldots, D\}$ be any set of k indexes. For each set of indexes T, we define a k-dimensional hyperplane in the D-dimensional input space: $\mathcal{X}_T \triangleq \{\mathbf{\Psi}\mathbf{z}, \text{ s.t. } \mathbf{z} \text{ is zero outside } T \text{ and } \|\mathbf{z}\| \leq 1\}$. By definition we have $\mathcal{X} = \cup_T \mathcal{X}_T$. We first show that Eqn 3 holds for each $\mathcal{X}_T$ and then use the union bound to prove the theorem.

For any given T, we choose a set $S \subset \mathcal{X}_T$ such that we have:

$$\forall \mathbf{x} \in \mathcal{X}_T : \min_{\mathbf{s}\in\mathbf{S}} \|\mathbf{x} - \mathbf{s}\| \leq \epsilon/4. \tag{4}$$

[1] The elements of the projection are typically taken to be distributed with $\mathcal{N}(0, 1/D)$, but we scale them by $\sqrt{D/d}$, so that we avoid scaling the projected values (see e.g. [10]).

It is easy to prove (see e.g. Chapter 13 of [13]) that these conditions can be satisfied by choosing a grid of size $|S| \leq (12/\epsilon)^k$, since $\mathcal{X}_T$ is a k-dimensional hyperplane in $\mathbb{R}^n$ (S fills up the space within $\epsilon/4$ distance). Now applying Theorem 1, and with $\|\mathbf{w}\| = 1$ we have that:

$$\forall \mathbf{s} \in \mathcal{S} : \left|\langle \mathbf{\Phi}^T \mathbf{w}, \mathbf{\Phi}^T \mathbf{s} \rangle - \langle \mathbf{w}, \mathbf{s} \rangle\right| \leq \frac{\epsilon}{2} \|\mathbf{s}\|, \tag{5}$$

fails with probability less than $(4(12/\epsilon)^k + 2)e^{-d\epsilon^2/192}$.

Let a be the smallest number such that:

$$\forall \mathbf{x} \in \mathcal{X}_T : \left|\langle \mathbf{\Phi}^T \mathbf{w}, \mathbf{\Phi}^T \mathbf{x} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle\right| \leq a \|\mathbf{x}\|, \tag{6}$$

holds when Eqn 5 holds. The goal is to show that $a \leq \epsilon$. For any given $\mathbf{x} \in \mathcal{X}_T$, we choose an $\mathbf{s} \in S$ for which $\|\mathbf{x} - \mathbf{s}\| \leq \epsilon/4$. Therefore we have:

$$\begin{aligned}
\left|\langle \mathbf{\Phi}^T \mathbf{w}, \mathbf{\Phi}^T \mathbf{x} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle\right| &\leq \left|\langle \mathbf{\Phi}^T \mathbf{w}, \mathbf{\Phi}^T \mathbf{x} \rangle - \langle \mathbf{\Phi}^T \mathbf{w}, \mathbf{\Phi}^T \mathbf{s} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle + \langle \mathbf{w}, \mathbf{s} \rangle\right| + && (7)\\
&\quad \left|\langle \mathbf{\Phi}^T \mathbf{w}, \mathbf{\Phi}^T \mathbf{s} \rangle - \langle \mathbf{w}, \mathbf{s} \rangle\right| && (8)\\
&\leq \left|\langle \mathbf{\Phi}^T \mathbf{w}, \mathbf{\Phi}^T (\mathbf{x} - \mathbf{s}) \rangle - \langle \mathbf{w}, (\mathbf{x} - \mathbf{s}) \rangle\right| + && (9)\\
&\quad \left|\langle \mathbf{\Phi}^T \mathbf{w}, \mathbf{\Phi}^T \mathbf{s} \rangle - \langle \mathbf{w}, \mathbf{s} \rangle\right| && (10)\\
&\leq a\epsilon/4 + \epsilon/2. && (11)
\end{aligned}$$

The last line is by the definition of a, and by applying Eqn 5 (with high probability). Because of the definition of a, there is an $\mathbf{x} \in \mathcal{X}_T$ (and by scaling, one with size 1), for which Eqn 6 is tight. Therefore we have $a \leq a\epsilon/4 + \epsilon/2$, which proves $a \leq \epsilon$ for any choice of $\epsilon < 1$.

Note that there are $\binom{D}{k}$ possible sets T. Since $\binom{D}{k} \leq (eD/k)^k$ and $\mathcal{X} = \cup_T \mathcal{X}_T$, the union bound gives us that the theorem fails with probability less than $(eD/k)^k(4(12/\epsilon)^k + 2)e^{-d\epsilon^2/192}$. □

4 Path-Wise LSTD

We consider the path-wise LSTD algorithm described in [14]. We observe a single trajectory $\{\mathbf{x}_t\}_{t=1}^n$ of size n generated by a Markov chain, along with observed rewards upon the transitions. We represent the input features by an $n \times D$ dimensional matrix $\mathbf{X}$ and the observed immediate rewards by a vector $\mathbf{r}$ of size n. Our goal is to estimate the value function, denoted by $v(\mathbf{x})$, which is defined to be the expected discounted sum of future rewards ($\sum_t \gamma^t r_t$), starting from any query state $\mathbf{x}$.

We define the shift operator $(\hat{P}\mathbf{y})_t \triangleq \mathbf{y}_{t+1}$ and the path-wise Bellman operator $\hat{B}\mathbf{y} \triangleq \mathbf{r} + \gamma \hat{P}\mathbf{y}$. For a linear space of approximators, there is a unique fixed point to the projected path-wise Bellman operator [14], which is the solution to a set of linear equations defined by:

$$\mathbf{X}^T(\mathbf{I} - \gamma \hat{P})\mathbf{X}\theta_{\text{lstd}} = \mathbf{X}^T \mathbf{r}. \tag{12}$$

We call θ_{lstd} the parameter of the LSTD solution and report $\hat{v}(\mathbf{x}) = \mathbf{x}^T \theta_{\text{lstd}}$ as our estimate of the value function for a query state $\mathbf{x}$. It is easy to see from Eqn 12, that the Bellman equation is satisfied on the observed trajectory: $\mathbf{X}\theta_{\text{lstd}} = \hat{B}\mathbf{X}\theta_{\text{lstd}}$.

5 Compressed Path-Wise LSTD

Instead of applying LSTD on the original space, one can first project the state features to a lower dimensional space using random projections and then solve for the fixed point to the projected path-wise Bellman operator in the new approximation space induced by the projection [9]. We use $\theta_{\text{lstd}}^{(\Phi)}$ to refer to the parameter of the LSTD solution after the projection $\mathbf{\Phi}$. We thus solve the set of equations defined by $(\mathbf{\Phi}^T \mathbf{X}^T)(\mathbf{I} - \gamma \hat{P})(\mathbf{X}\mathbf{\Phi})\theta_{\text{lstd}}^{(\Phi)} = (\mathbf{\Phi}^T \mathbf{X}^T)\mathbf{r}$, and estimate the value function as $\hat{v}(\mathbf{x}) = (\mathbf{x}^T \mathbf{\Phi})\theta_{\text{lstd}}^{(\Phi)}$.

To get a bound on the accuracy of this estimate, we make two assumptions. First, we assume that the state features are in a k-sparse, D-dimensional and norm-bounded space: $\mathbf{x}_t \in \mathcal{X}$. Second, we

assume that the value function is almost linear with a parameter θ and an additive bias bounded by some $\epsilon_v \geq 0$:

$$\forall \mathbf{x} \in \mathcal{X} : |v(\mathbf{x}) - \mathbf{x}^T\theta| \leq \epsilon_v. \tag{13}$$

Applying Theorem 2 to the above assumption, we can see that the value function is almost-linear in the projected space with parameter $\mathbf{\Phi}^T\theta$:

Corollary 3. *Let $\mathbf{\Phi}^{D\times d}$ be a random projection according to Eqn 1. Assume that the value function $v(.)$ is almost linear in the original features with parameter θ and an additive bias bounded by some $\epsilon_v \geq 0$, as defined in Eqn 13. Then for any $0 < \delta_{prj} < 1$ with probability no less than $1 - \delta_{prj}$:*

$$\forall \mathbf{x} \in \mathcal{X} : \left|v(\mathbf{x}) - (\mathbf{x}^T\mathbf{\Phi})(\mathbf{\Phi}^T\theta)\right| \leq \epsilon_v + \epsilon_{prj}\|\theta\|, \tag{14}$$

where,

$$\epsilon_{prj} = c\sqrt{(k\log d/d)\log(12eD/k\delta_{prj})}. \tag{15}$$

The above corollary suggests that we can use LSTD after the projection, as the value function can be closely approximated in the projected space. We are now ready to describe our main theorem on the worst-case error in our estimated values (proof included in the appendix):

Theorem 4. *Let $\mathbf{\Phi}^{D\times d}$ be a random projection according to Eqn 1 and $\theta_{lstd}^{(\Phi)}$ be parameter to the path-wise LSTD solution in the compressed space induced by the projection. Assume that the value function $v(.)$ is almost linear in the original features with parameter θ and an additive bias bounded by some $\epsilon_v \leq 0$, as defined in Eqn 13. Also, assume that the joint distribution $P(\mathbf{x}_1, ..., \mathbf{x}_t)$ for all $t > 1$ is absolutely continuous w.r.t. the appropriate Lebesgue measure. Let $\theta_Z = \mathbf{X}^\dagger\mathbf{v}$ where $\mathbf{v}$ is the true value of the observed states. Choose any $0 < \delta_{prj}, \delta_{emp}, \delta_{var} < 1$ and $d \geq 15\log(8n/\delta_{emp})$. Then, with probability no less than $1 - (\delta_{var} + \delta_{prj} + \delta_{emp})$, we have $\forall \mathbf{x} \in \mathcal{X}$:*

$$\begin{aligned}\left|\mathbf{x}^T\mathbf{\Phi}\theta_{lstd}^{(\Phi)} - v(\mathbf{x})\right| \quad \leq \quad & \left(\epsilon_v + \epsilon_{prj}\|\theta\|\right)(1+\rho\sqrt{n}) + \\ & \gamma V_{\max}\rho^2\sqrt{8n\log(2d/\delta_{var})} + \\ & \frac{\gamma\rho}{\sqrt{1-\gamma^2}}\left(\|\mathbf{v} - \mathbf{X}\theta_Z\| + \|\theta_Z\|\sqrt{\frac{8n\log(8n/\delta_{emp})}{d}}\right) + \\ & \frac{\gamma^2\rho V_{\max}}{(1-\gamma)\sigma_{\min}^{(X)}}\left(\sqrt{8n\log(4d/\delta_{emp})} + 1\right),\end{aligned} \tag{16}$$

where $\rho = (1+\epsilon_\Phi)\|(\mathbf{X}\mathbf{\Phi})^\dagger\|$, with $\epsilon_\Phi = \sqrt{(12/d)log(2/\delta_{prj})}$ and ϵ_{prj} is as defined in Eqn 15.

Notice that because we use random projections of the type defined in Eqn 1, the norm of $\mathbf{\Phi}$ can be bounded using the bound discussed in [15]; we have with probability $1 - \delta_\Phi$:

$$\|\mathbf{\Phi}\| \leq \sqrt{D/d} + \sqrt{(2\log(2/\delta_\Phi))/d} + 1 \quad \text{and} \quad \|\mathbf{\Phi}^\dagger\| \leq \left[\sqrt{D/d} - \sqrt{(2\log(2/\delta_\Phi))/d} - 1\right]^{-1}.$$

When $n > D$, and the observed states are sufficiently spread out, then $\|\mathbf{X}^\dagger\|$ is of order $\tilde{O}(\sqrt{D/n})$ and thus $\rho = \tilde{O}(\sqrt{d/n})$ (under similar assumptions, this is also true when $d < n < D$, but needs further anlysis). We substitute this into the error bound of Theorem 4 and ignore the logarithmic terms and those independent of d and n, and assume $\epsilon_v = 0$. We observe that the first term, of order $\tilde{O}(\sqrt{k/d})$, is a bias term due to the projection that decreases by increasing d. The second term, of order $\tilde{O}(d/\sqrt{n})$, is a variance term that increases with d. The third and forth terms are of the same order as the on-sample bound used in [9], which includes a term of order $\tilde{O}(\sqrt{d})$. We get:

$$\left|\mathbf{x}^T\mathbf{\Phi}\theta_{\text{lstd}}^{(\Phi)} - v(\mathbf{x})\right| \leq \tilde{O}(\sqrt{k}/\sqrt{d}) + \tilde{O}(d/\sqrt{n}) + \tilde{O}(\sqrt{d}). \tag{17}$$

A more careful analysis of this bound is a subject of future work. We conjecture that the second term can be tightened by a $1/\sqrt{d}$ factor, using a bound similar to that of Lemma 5 that holds with high probability for a fixed $\mathbf{x} \in \mathcal{X}$, and the last term can also be tightened by a $1/\sqrt{n}$ factor using the same mechanism. Therefore we propose the following bound. With probability no less than $1 - \delta_{\text{prj}}$, for any fixed $\mathbf{x} \in \mathcal{X}$, with probability no less than $1 - (\delta_{\text{var}} + \delta_{emp})$:

$$\left|\mathbf{x}^T\mathbf{\Phi}\theta_{\text{lstd}}^{(\Phi)} - v(\mathbf{x})\right| \leq \tilde{O}(\sqrt{k}/\sqrt{d}) + \tilde{O}(\sqrt{d}/\sqrt{n}), \tag{18}$$

which suggests an optimal projection size of $d = \sqrt{nk}$.

References

[1] A.M. Farahmand, M. Ghavamzadeh, C. Szepesvári, and S. Mannor. Regularized fitted Q-iteration for planning in continuous-space Markovian decision problems. In *American Control Conference (ACC)*, 2009.

[2] A.M. Farahmand, M. Ghavamzadeh, Cs. Szepesvári, and S. Mannor. Regularized policy iteration. In *Neural Information Processing Systems (NIPS)*, 2009.

[3] J.Z. Kolter and A.Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In *International Conference on Machine Learning (ICML)*, 2009.

[4] I. Menache, S. Mannor, and N. Shimkin. Basis function adaptation in temporal difference reinforcement learning. *Annals of Operations Research*, 134(1):215–238, 2005.

[5] P.W. Keller, S. Mannor, and D. Precup. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2006.

[6] R. Parr, C. Painter-Wakefield, L. Li, and M. Littman. Analyzing feature generation for value-function approximation. In *International Conference on Machine Learning (ICML)*, 2007.

[7] D.L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[8] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

[9] M. Ghavamzadeh, A. Lazaric, O.A. Maillard, and R. Munos. LSTD with random projections. In *Neural Information Processing Systems (NIPS)*, 2010.

[10] M.A. Davenport, M.B. Wakin, and R.G. Baraniuk. Detection and estimation with compressive measurements. *Dept. of ECE, Rice University, Tech. Rep*, 2006.

[11] E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.

[12] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. The Johnson–Lindenstrauss lemma meets compressed sensing. *Constructive Approximation*, 2007.

[13] G.G. Lorentz, M. von Golitschek, and Y. Makovoz. *Constructive approximation: advanced problems*, volume 304. Springer Berlin, 1996.

[14] A. Lazaric, M. Ghavamzadeh, and R. Munos. Finite-Sample Analysis of LSTD. In *Proceedings of the international conference on machine learning*, 2010.

[15] E.J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies. *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

Appendix

Recall the notion of regression with Markov design, introduced in [14]:

Definition 1. *The model of regression with **Markov design** is a regression problem where the data $(\mathbf{x}_t, Y_t)_{1\le t\le n}$ is generated according to the following model: $\mathbf{x}_1, ..., \mathbf{x}_t \in \mathbb{R}^D$ is a sample path generated by a Markov chain, $Y_t = f(\mathbf{x}_t) + \eta_t$, where f is the target function, and the noise term η_t is a random variable which is adapted to the filtration generated by $\mathbf{x}_1, ..., \mathbf{x}_{t+1}$ and is such that*

$$|\eta_t| \le C, \quad \text{and} \quad E(\eta_t|\mathbf{x}_1, ..., \mathbf{x}_t) = 0.$$

The following lemma provides a worst case bound of a regression under the Markov design setting in a projected space. As we have mentioned, we assume that the norm of an original signal $\mathbf{x} \in \mathcal{X}$, is bounded by 1. Since we are interested in applying a worst case bound in the projected space, we will be working with a new (projected) signal $\mathbf{\Phi}^T\mathbf{x}$, which is bounded by $\|\mathbf{\Phi}\|$ for a specific choice of $\mathbf{\Phi}$.

Lemma 5. *Let $\mathbf{\Phi} \in \mathbb{R}^{D\times d}$ be a full rank matrix. In a Markov design setting w.r.t. $(\mathbf{x}_t \in \mathcal{X}, Y_t)_{1\le t\le n}$, let $\hat\theta \in \mathbb{R}^d$ be the parameter found by the OLS on a projected data $\mathbf{X\Phi}$ and (noisy) output $\mathbf{y} = \{Y_t\}_1^n$; and let $\theta \in \mathbb{R}^d$ be the parameter found by the OLS on the same projected data $\mathbf{X\Phi}$ and output $f(\mathbf{X}) = \{f(\mathbf{x}_t)\}_1^n$. Assume that the joint distribution $P(\mathbf{x}_1, ..., \mathbf{x}_t)$ for all $t > 1$ is absolutely continuous w.r.t. the appropriate Lebesgue measure. Then, for any $1 > \delta > 0$, with probability $1 - \delta$ w.r.t. the sample path $\mathbf{x}_1, ..., \mathbf{x}_n$, the noise η and (future) signal $\mathbf{x}$ we have*

$$\|(\mathbf{\Phi}^T\mathbf{x})^T\theta - (\mathbf{\Phi}^T\mathbf{x})^T\hat\theta\| \le C\|\mathbf{x}_m^T\mathbf{\Phi}\|^2\|(\mathbf{X\Phi})^\dagger\|^2\sqrt{2n\log(2d/\delta)},$$

where $\mathbf{x}_m = \arg\max_{\mathbf{z}\in\mathcal{X}} \|\mathbf{z}^T\mathbf{\Phi}\|$.

Proof. We have:

$$\begin{aligned}
\|(\mathbf{\Phi}^T\mathbf{x})^T\theta - (\mathbf{\Phi}^T\mathbf{x})^T\hat\theta\| &\le \|\mathbf{x}_m^T\mathbf{\Phi}\|\|\theta - \hat\theta\| \\
&\le \|\mathbf{x}_m^T\mathbf{\Phi}\|\|(\mathbf{X\Phi})^\dagger f(\mathbf{X}) - (\mathbf{X\Phi})^\dagger Y\| \qquad (19)\\
&= \|\mathbf{x}_m^T\mathbf{\Phi}\|\|(\mathbf{X\Phi})^\dagger\eta\| \\
&= \|\mathbf{x}_m^T\mathbf{\Phi}\|\|[(\mathbf{X\Phi})^T\mathbf{X\Phi}]^{-1}(\mathbf{X\Phi})^T\eta\| \qquad (20)\\
&\le \|\mathbf{x}_m^T\mathbf{\Phi}\|\|(\mathbf{\Phi}^T\mathbf{X}^T\mathbf{X\Phi})^{-1}\|\|(\mathbf{X\Phi})^T\eta\| \\
&\le \|\mathbf{x}_m^T\mathbf{\Phi}\|\|(\mathbf{X\Phi})^\dagger\|^2\|(\mathbf{X\Phi})^T\eta\|. \qquad (21)
\end{aligned}$$

Note that Equation 20 holds since $\mathbf{X\Phi}$ is full rank with probability 1 due the assumptions on $\mathbf{X}$ and $\mathbf{\Phi}$. Now, observe that for any fixed $\mathbf{\Phi}$,

$$E(\eta_n|\mathbf{\Phi}^T\mathbf{x}_1, ..., \mathbf{\Phi}^T\mathbf{x}_n) = E\Big[E(\eta_n|\mathbf{x}_1, ..., \mathbf{x}_n)\Big|\mathbf{\Phi}^T\mathbf{x}_1, ..., \mathbf{\Phi}^T\mathbf{x}_n\Big] = 0,$$

due to Markov design setting. Therefore, for each $d \ge i \ge 1, n \ge j \ge 1$:

$$E(\Phi_i^T\mathbf{x}_j\eta_j|\mathbf{\Phi}^T x_1, ..., \mathbf{\Phi}^T\mathbf{x}_j) = \Phi_i^T\mathbf{x}_j E(\eta_j|S_1, ..., S_j) = 0,$$

where Φ_i is the i-th column of $\mathbf{\Phi}$. Hence, similarly to the proof of Lemma 1 in [14], $\Phi_i^T\mathbf{x}_j\eta_j$ are martingale differences, so we can apply Azuma's inequality with the union bound over all $d \ge i \ge 1$, and obtain: for any $1 > \delta > 0$ we have with probability $1 - \delta$:

$$\left|\sum_{j=1}^n \Phi_i^T\mathbf{x}_j\eta_j\right| \le C\sqrt{2\left(\sum_{j=1}^n[\Phi_i^T\mathbf{x}_j]^2\right)\log(2d/\delta)},$$

from which we get:

$$\begin{aligned}
\|\mathbf{\Phi}^T\mathbf{X}^T\eta\| &= \sqrt{\sum_{i=1}^d\Big(\sum_{j=1}^n\Phi_i^T\mathbf{x}_j\eta_j\Big)^2} \\
&\le C\sqrt{2\log(2d/\delta)\sum_{j=1}^n\sum_{i=1}^d[\Phi_i^T\mathbf{x}_j]^2} = C\sqrt{2\log(2d/\delta)\sum_{j=1}^n\|\mathbf{\Phi}^T\mathbf{x}_j\|^2} \\
&\le C\|\mathbf{x}_m^T\mathbf{\Phi}\|\sqrt{2n\log(2d/\delta)},
\end{aligned}$$

completing the proof. □

Proof of Theorem 4

Proof. Let $\theta_Z^{(\Phi)}$ be the OLS solution for the regression of $\mathbf{v}$ on $\mathbf{X}\mathbf{\Phi}$, and $\theta_Y^{(\Phi)}$ be the OLS solution for the regression of $\hat{B}\mathbf{v}$ on $\mathbf{X}\mathbf{\Phi}$. Again, we define $\mathbf{x}_m = \arg\max_{\mathbf{z}\in\mathcal{X}} \|\mathbf{z}^T\mathbf{\Phi}\|$.

We have that for $\forall \mathbf{x} \in \mathcal{X}$:

$$\begin{aligned}\left|\mathbf{x}^T\mathbf{\Phi}\theta_{\text{lstd}}^{(\Phi)} - v(\mathbf{x})\right| \;\leq\; &\left|\mathbf{x}^T\mathbf{\Phi}\theta_Z^{(\Phi)} - v(\mathbf{x})\right| + \\ &\left|\mathbf{x}^T\mathbf{\Phi}\theta_Y^{(\Phi)} - \mathbf{x}^T\mathbf{\Phi}\theta_Z^{(\Phi)}\right| + \\ &\left|\mathbf{x}^T\mathbf{\Phi}\theta_{\text{lstd}}^{(\Phi)} - \mathbf{x}^T\mathbf{\Phi}\theta_Y^{(\Phi)}\right|.\end{aligned} \tag{22}$$

The first term of the above is a bias term. Let $\mathbf{b} = \mathbf{v} - \mathbf{X}\mathbf{\Phi}\mathbf{\Phi}^T\theta$. Using Corollary 3, we have that $\|\mathbf{b}\| \leq (\epsilon_v + \epsilon_{\text{prj}}\|\theta\|)\sqrt{n}$. Using the definition of $\theta_Z^{(\Phi)}$ we get:

$$\begin{aligned}\left|\mathbf{x}^T\mathbf{\Phi}\theta_Z^{(\Phi)} - v(\mathbf{x})\right| &\leq \left|\mathbf{x}^T\mathbf{\Phi}(\mathbf{X}\mathbf{\Phi})^\dagger\mathbf{v} - \mathbf{x}^T\mathbf{\Phi}\mathbf{\Phi}^T\theta\right| + \left|\mathbf{x}^T\mathbf{\Phi}\mathbf{\Phi}^T\theta - v(\mathbf{x})\right| && (23)\\ &\leq \left|\mathbf{x}^T\mathbf{\Phi}(\mathbf{X}\mathbf{\Phi})^\dagger(\mathbf{X}\mathbf{\Phi}\mathbf{\Phi}^T\theta + \mathbf{b}) - \mathbf{x}^T\mathbf{\Phi}\mathbf{\Phi}^T\theta\right| + \epsilon_v + \epsilon_{\text{prj}}\|\theta\| && (24)\\ &\leq \left|\mathbf{x}^T\mathbf{\Phi}(\mathbf{X}\mathbf{\Phi})^\dagger\mathbf{b}\right| + \epsilon_v + \epsilon_{\text{prj}}\|\theta\| && (25)\\ &\leq \|\mathbf{x}^T\mathbf{\Phi}\|\|(\mathbf{X}\mathbf{\Phi})^\dagger\|\|\mathbf{b}\| + \epsilon_v + \epsilon_{\text{prj}}\|\theta\| && (26)\\ &\leq \left(\epsilon_v + \epsilon_{\text{prj}}\|\theta\|\right)\left(1 + \sqrt{n}\|\mathbf{x}_m^T\mathbf{\Phi}\|\|(\mathbf{X}\mathbf{\Phi})^\dagger\|\right). && (27)\end{aligned}$$

The second term in Eqn 22 is a variance term. Using Lemma 5, taking $C = 2\gamma V_{\max}$, with probability no less than $1 - \delta_{\text{var}}$ we can bound this term

$$\left|\mathbf{x}^T\mathbf{\Phi}\theta_Y^{(\Phi)} - \mathbf{x}^T\mathbf{\Phi}\theta_Z^{(\Phi)}\right| \leq 2\gamma V_{\max}\|\mathbf{x}_m^T\mathbf{\Phi}\|^2\|(\mathbf{X}\mathbf{\Phi})^\dagger\|^2\sqrt{2n\log(2d/\delta_{\text{var}})}. \tag{28}$$

To bound the third term of Eqn 22, consider the Bellman equation $(\mathbf{X}\mathbf{\Phi})\theta_{\text{lstd}}^{(\Phi)} = \hat{B}(\mathbf{X}\mathbf{\Phi})\theta_{\text{lstd}}^{(\Phi)}$, from which (since $\mathbf{X}\mathbf{\Phi}$ is full-rank w.p. 1 due to the assumptions on $\mathbf{X}$ and $\mathbf{\Phi}$) we conclude that $\theta_{\text{lstd}}^{(\Phi)} = (\mathbf{X}\mathbf{\Phi})^\dagger\hat{B}(\mathbf{X}\mathbf{\Phi})\theta_{\text{lstd}}^{(\Phi)}$. By definition we have that $\theta_Y^{(\Phi)} = (\mathbf{X}\mathbf{\Phi})^\dagger\hat{B}\mathbf{v}$. Therefore, for the third term of Eqn 22 we have:

$$\begin{aligned}\left|\mathbf{x}^T\mathbf{\Phi}\theta_{\text{lstd}}^{(\Phi)} - \mathbf{x}^T\mathbf{\Phi}\theta_Y^{(\Phi)}\right| &\leq \left|\mathbf{x}^T\mathbf{\Phi}(\mathbf{X}\mathbf{\Phi})^\dagger\hat{B}(\mathbf{X}\mathbf{\Phi})\theta_{\text{lstd}}^{(\Phi)} - \mathbf{x}^T\mathbf{\Phi}(\mathbf{X}\mathbf{\Phi})^\dagger\hat{B}\mathbf{v}\right| && (29)\\ &\leq \left\|\mathbf{x}^T\mathbf{\Phi}\right\|\left\|(\mathbf{X}\mathbf{\Phi})^\dagger\right\|\left\|\hat{B}(\mathbf{X}\mathbf{\Phi})\theta_{\text{lstd}}^{(\Phi)} - \hat{B}\mathbf{v}\right\| && (30)\\ &\leq \gamma\left\|\mathbf{x}_m^T\mathbf{\Phi}\right\|\left\|(\mathbf{X}\mathbf{\Phi})^\dagger\right\|\left\|(\mathbf{X}\mathbf{\Phi})\theta_{\text{lstd}}^{(\Phi)} - \mathbf{v}\right\|. && (31)\end{aligned}$$

Line 31 is by the contraction of the $\hat{B}$ operator. For the last term on the RHS, we apply an adaptation of Theorem 1 from [9]:

$$\begin{aligned}\left\|(\mathbf{X}\mathbf{\Phi})\theta_{\text{lstd}}^{(\Phi)} - \mathbf{v}\right\| \;\leq\; &\frac{1}{\sqrt{1-\gamma^2}}\left(\|\mathbf{v} - \mathbf{X}\theta_Z\| + \|\theta_Z\|\sqrt{\frac{8n\log(8n/\delta_{\text{emp}})}{d}}\right) + \\ &\frac{\gamma V_{\max}}{(1-\gamma)\sigma_{\min}^{(X)}}\left(\sqrt{8n\log(4d/\delta_{\text{emp}})} + 1\right).\end{aligned} \tag{32}$$

Combining the three terms and using the upper bound $\left\|\mathbf{x}_m^T\mathbf{\Phi}\right\| \leq (1+\epsilon_\Phi)\left\|\mathbf{x}_m^T\right\| \leq (1+\epsilon_\Phi)$ (used within the proof of Theorems 1 and 2), give us the bound in the theorem.

□