Compressed Least-Squares Regression on Sparse Spaces

Mahdi Milani Fard, Yuri Grinberg

Doina Precup, Joelle Pineau

Reasoning and Learning Laboratory

School of Computer Science

Sparse Spaces

Features are in a k-sparse compact subspace of $\mathbb{R}^D$:

$$\mathcal{X} \triangleq \{\Psi \mathbf{z}, \text{ s.t. } \|\mathbf{z}\|_0 \leq k \text{ and } \|\mathbf{z}\| \leq 1\}$$

$$0\,0 \ldots 0\,0\,3\,0\,0 \ldots 0\,0\,1\,0\,0 \ldots 0\,0\,5\,0\,0 \ldots 0\,0$$

Images, video data, music, audio …

Discretized spaces, tile-coding, …

Random Projections

Use random projections to compress the space

Linear operator $\Phi^{D \times d}$ where $\Phi_{i,j} \sim \mathcal{N}(0, 1/d)$

Random Projections in Sparse Spaces

$\mathcal{X}$ is D dimensional k-sparse space

projection size $d = \tilde{O}(k \log D)$

$$\Downarrow$$

$$\forall \mathbf{x} \in \mathcal{X} : \|\mathbf{x}\| \simeq \|\Phi^T \mathbf{x}\|$$

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{X} : \|\mathbf{x} - \mathbf{y}\| \simeq \|\Phi^T \mathbf{x} - \Phi^T \mathbf{y}\|$$

(E.g. Achlioptas, 2001)

Random Projections and Sparse Spaces

Compression

- With high prob. can recover $\mathbf{x}$ from $\Phi^T \mathbf{x}$
- Requires non-linear optimization

Detection/Classification

- With high prob. can detect if $\mathbf{x} = \mathbf{y}$
- Need only to compare $\Phi^T \mathbf{x}$ and $\Phi^T \mathbf{y}$

(E.g. Davenport et al., 2010)

Regression?

Linearity in Sparse Spaces

Random projections of size $O(k \log D)$ preserve linearity for sparse spaces.

For any fixed $\mathbf{w} \in \mathbb{R}^D$ and any k-sparse space $\mathcal{X}$:

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

fails with probability $< e^{O(\log(D) - d\epsilon^2/k)}$

Compressed Ordinary Least-Squares

Input: Training set $(\mathbf{X}, \mathbf{y})$, projection size d

Output: $\hat{\mathbf{w}} \in \mathbb{R}^D$, linear coefficient of the approximator

Apply OLS in the compressed space: $\mathbf{w}_{\text{ols}}^{(\Phi)} = (\mathbf{X}\Phi)^{\dagger}\mathbf{y}$;

Output $\hat{\mathbf{w}} \leftarrow \Phi\mathbf{w}_{\text{ols}}^{(\Phi)}$;

Finite Sample Analysis

i.i.d training set + Gaussian noise

$$y = \mathbf{x}^T \mathbf{w} + \eta$$

Well distributed sample

$$\left| \mathbf{x}^T \mathbf{w} - \mathbf{x}^T \hat{\mathbf{w}} \right| \leq \tilde{O}\left(\sqrt{k \log(D/k)} \frac{1}{\sqrt{d}} \right) + \tilde{O}\left(\frac{\sigma_\eta}{\sqrt{n}} \sqrt{d} \right)$$

Finite Sample Analysis

Optimal projection size $\tilde{O}(\sqrt{kn \log D})$

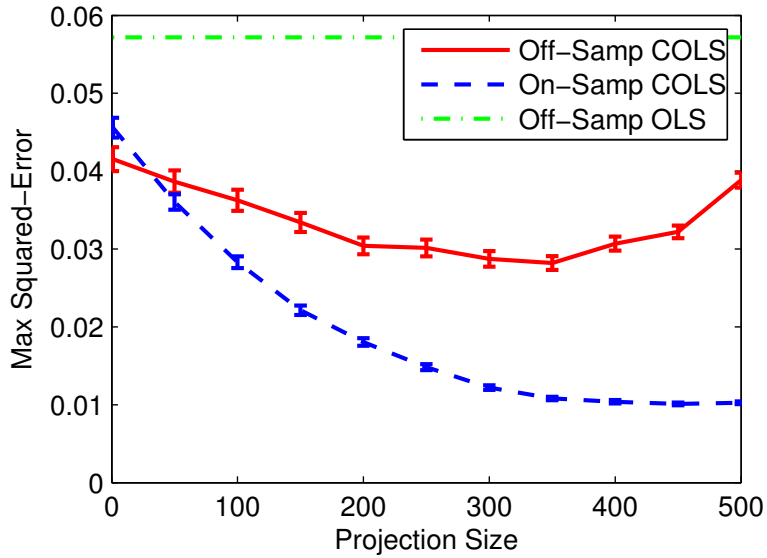
Resembles $\tilde{O}(\sqrt{n \log D})$ for general non-sparse spaces
(Maillard and Munos, 2009)

Empirical Results

Synthetic domain:

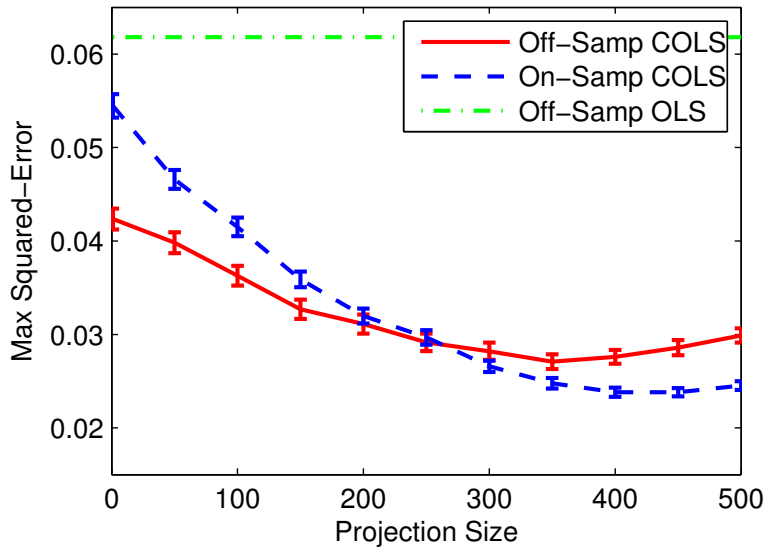
- $D = 1000$
- $k = 50$ (5% non-zero)
- Target $\mathbf{w}$ generated randomly

Empirical Results



$n = 800 < D$, SNR $\simeq 1$.

Empirical Results

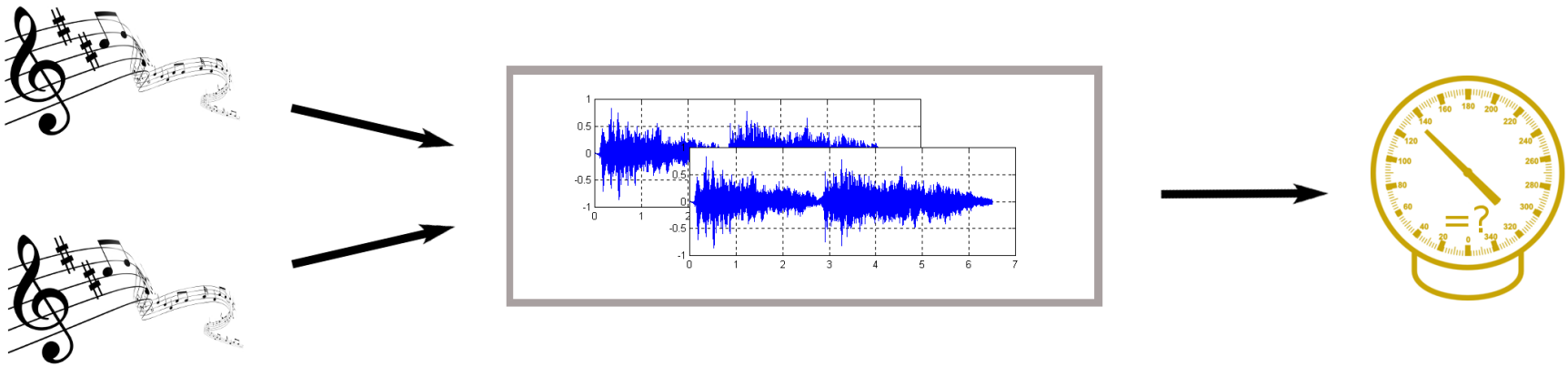


$n = 2000 > D$, SNR $\simeq 0.4$.

Empirical Results

Similarity measure between music tracks based on playlists and listening patterns

Empirical Results



Task: predict similarity based on audio analysis

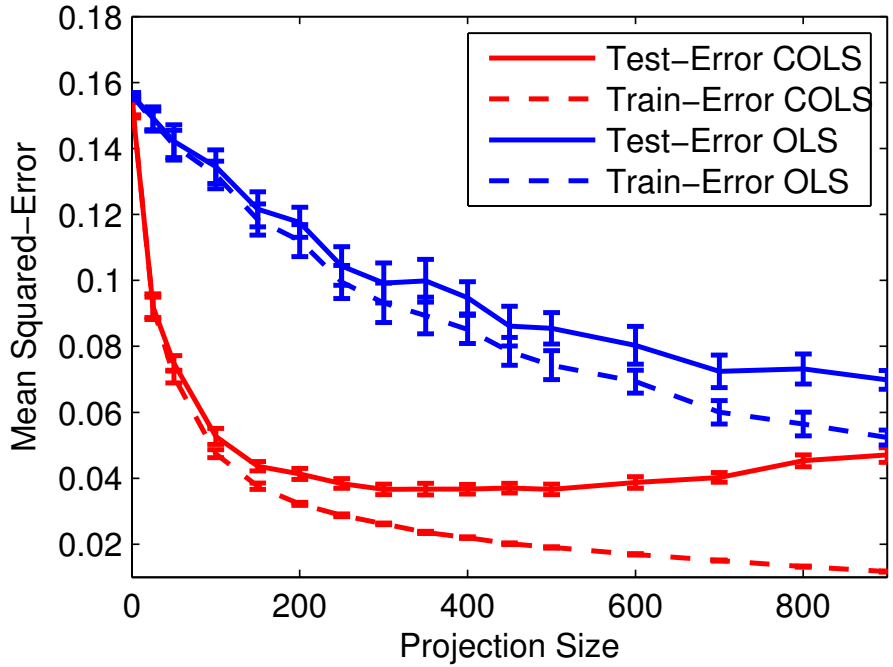
Music similarity prediciton

Empirical Results

Music similarity prediction:

- Features: co-presence of common chord-progressions
- 10^6 features, very sparse
- 2000 samples

Sparsity of Linear Coefficient



COLS vs. OLS with randomly chosen features

Discussion

Take home message:

- Random projections can be used for regression
- Effective linear dimensionality of sparse spaces is logarithmic in the nominal dimensionality

Future work:

- Different noise model
- Consider the effect of sparsity in the linear coefficient
- Regularized regression (e.g. ridge regression)

Questions?

References:

- D. Achlioptas, Database-friendly Random Projections. PODS 2001
- M.A. Davenport, M.B. Wakin, and R.G. Baraniuk. Detection and Estimation with Compressive Measurements. Tech. Rep, 2006.
- O.A. Maillard and R. Munos. Compressed Least-squares Regression. NIPS, 2009.
- M. Davenport, P. Boufounos, M. Wakin, and R. Baraniuk. Signal processing with compressive measurements. IEEE J. Select. Top. Signal Processing, 4(2): 445–460, 2010
- M.M. Fard, Y. Grinberg, J. Pineau, D. Precup. Bellman Error Based Feature Generation Using Random Projections. EWRL, 2012.

Theorem

Let $\mathbf{w}_{ols}^{(\Phi)}$ *be the OLS solution in the compressed space induced by the projection. Assume an additive bias in the original space bounded by some* $\epsilon_f > 0$ *and i.i.d. noise with variance* σ_η^2*. Choose any* $0 < \delta_{prj} < 1$ *and* $0 < \delta_{var} \leq \sqrt{2/e\pi}$ *. Then, with probability no less than* $1 - \delta_{prj}$*, we have for any* $\mathbf{x} \in \mathcal{X}$ *with probability no less than* $1 - \delta_{var}$*:*

$$
\begin{aligned}
&|f(\mathbf{x}) - \mathbf{x}^T \Phi \mathbf{w}_{ols}^{(\Phi)}| \leq \\
&\qquad \|\mathbf{x}^T \Phi\| \|(\mathbf{X}\Phi)^\dagger\| \left(\left(\epsilon_f + \epsilon_{prj}\right) \sqrt{n} + \sigma_\eta \sqrt{\log(2/\pi\delta_{var}^2)} \right) \\
&\qquad + \epsilon_f + \epsilon_{prj},
\end{aligned}
$$

where $\epsilon_{prj} = \sqrt{\frac{48k}{d} \log \frac{4D}{\delta_{prj}}}$.