Grady Feature Selection for Supervised Learning

- matching pursuit (Mallet & Zhang 1993)
  - use correlation between residual & candidate features
  - select feature w/highest correlation

- orthogonal matching pursuit for regression (OMP)
  - recomputes residual after each feature added

- OMP algorithm
  - input: \( X \in \mathbb{R}^{n \times k} \), \( y \in \mathbb{R}^n \), \( \beta \in \mathbb{R} \) (given input feature set)
  - output: approx weights \( w \)
  - \( I = \emptyset \)
  - \( w = 0 \) (0 vector, length \( k \))
  - repeat
    - \( c_i = |X^T (y - Xw)| \)
    - \( j = \arg \max_i c_i \)
    - if \( c_j > \beta \), add \( j \) to \( I \)
    - \( w_j = X^+_j \cdot y \)
  - until \( c_j \leq \beta \) or \( I = \{ \} \) (all features added)

- comp cost: at least \( n \times k \) per iteration
  - \( O(k^2) \) for inversion

- OMP properties

  define \( y \) as \( m \)-sparse in \( X \) if \( \exists X_{opt} \) composed of \( m \) columns of \( X \) s.t. \( y = X_{opt} \cdot w_{opt} \)
  - and there is no \( X' \) w/fewer columns of \( X \) s.t. \( y \)
    can be represented as \( y = X' \cdot w \)

  thm (Tropp 2004)
  - if \( y \) is \( m \)-sparse in \( X \) and \( (1) \)
    - \( \max_{i \notin opt} \|X_{opt}^T \cdot x_i\|_1 < 1 \) \( (2) \)
  - then OMP called with \( X, y \) and \( \beta = 0 \) will return \( w_{opt} \)
    in \( m \) iterations.

  note: any orthogonal basis \( X \) satisfies equation 2.
  - the above (I believe) holds for \( \infty \) data, exact sparsity
    - no noise
  - Tropp has an extension for approx sparsity
  - Zhang (2009) has extension for noisy case
  - but still assumes infinite data
Recall LSTD fixed point & Bellman residual minimization

Bellman residual projection $\pi$

Fixed pt error $V$

$\pi V$

Bellman residual min $w$ w.r.t.

$$\min_w \| R + \gamma \bar{V} w - \bar{V} w \|^2 = \| R - (\bar{V'} - \gamma \bar{V'}) w \|^2$$

Use OMP algorithm where

$y = R$

$X = \bar{V'} - \gamma \bar{V'}$

OMP-TD (OMP w/LSTD)

Interleave fixed pt calculation given current selected features with adding features that correlate most with Bellman error

OMP-TD algorithm

Input

$\bar{V} \in \mathbb{R}^{n \times k}$: $\bar{V}^i_j = y_j(s_i)$

$\bar{V}' \in \mathbb{R}^{n \times k}$: $\bar{V}'^i_j = y_j(s'_i)$

$R \in \mathbb{R}^n$, $R_i = r_i$

$q \in [0,1]$, $\beta \in \mathbb{R}$

Output: $w$

$I \in \{1\}$

$w = 0$

Repeat

$c = \| R + \gamma \bar{V} w - \bar{V} w \|^2 / n$ \hspace{1em} // find features correlate $w$ with Bellman error

$j = \arg \max_i i \in X \cap c$

if $c_j > \beta$, $X = X \cup j$

$w_X = (\bar{V}' \bar{V}^{-1} \bar{V} - \gamma \bar{V}' \bar{V}^{-1} \bar{V})^{-1} \bar{V} R$ \hspace{1em} // compute fixed point using selected features

Until $c_j \leq \beta$ or $X = \emptyset$

Guarantees/Properties

OMPTD

Good news: if feature j adding has a corresponding $c_j >$ some threshold

Then it improves a bound on the distance between $V^*$ and the fixed point
but can be suboptimal consider
\[ S_i \rightarrow S_{i+1} \]
rewards:
\[
V^* = \begin{bmatrix}
-\gamma \cdot y^2 & 1 & 1 & 1 & 1 \\
0 & 1 & -\gamma \cdot y^2 & 1 & 0
\end{bmatrix}
\]
let \( \mathbf{v} \) be an orthonormal basis defined by indicator functions
\[ \psi_i(s) = 1_{s=s_i} \]
\( V^* \) is 3-sparse in \( \mathbf{v} \)
\( \mathbf{v} \) opt only requires features 2, 3 & 4

exercise
at start \( w = 0 \)
1) what is residual vector? \( (R + \gamma \cdot \mathbf{v} \cdot \omega - \mathbf{v} \cdot \omega) \)
assume have \( \mathbf{v} \) and \( \mathbf{v}' \) for all states
2) what feature will OMP-TD select first?
\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
-\gamma \cdot y^2 & y^2 & 1 & 0 & 0
\end{bmatrix}
= \begin{bmatrix}
1 & 1 & 0 & 0 & 0
\end{bmatrix}
\]
\( \gamma = 0.65 \), 1st feature

problem: implies even if \( V^* \) is m-sparse, OMP-TD cannot guarantee recovery in m iterations
eg. may include extra features
can modify example to make many unnecessary features added

what about OMP-BRM?
lemma 1: if \( V^* \) is m-sparse in \( \mathbf{v} \) then \( R \) is m-sparse in \( (\mathbf{v} - \gamma P \mathbf{v}) \)

proof: recall \( V^* = (I - \gamma P)^{-1} R \)
\[
(I - \gamma P) R = \mathbf{v}_{opt} w_{opt}
\]
\[
R = \mathbf{v}_{opt} w_{opt} - \gamma P \mathbf{v}_{opt} w_{opt}
\]
\[
= (I_{opt} - \gamma P \mathbf{v}_{opt}) w_{opt}
\]
assume other direction, that have found a m-sparse rep of \( R \) with param \( w_{opt} \) in basis \( (\mathbf{v} - \gamma P \mathbf{v}_{opt}) \)
then \( R = (I_{opt} - \gamma P \mathbf{v}_{opt}) w_{opt} \)
\[
= (I - \gamma P) \mathbf{v}_{opt} \]
\[
(I - \gamma P)^{-1} R = \mathbf{v}_{opt} w_{opt}
\]
\[
V = \mathbf{v}_{opt} w_{opt}
\]
\( \Rightarrow \) if perform OMP on \( (\mathbf{v} - \gamma P \mathbf{v}) \) basis on target \( R \), then indices of \( (\mathbf{v} - \gamma P \mathbf{v}) \) should be selected to rep \( V \) (and \( W = \mathbf{v}_{opt} \) same)
exercise: some example, run OMP-TD

1) What’s $X$?

$X = \frac{\bar{\mathbf{z}}}{-\gamma \bar{\mathbf{p}}}$

$X \in \mathbb{R}^{n \times k}$

$X(s_1,:) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & -\gamma & 0 & 0 \end{bmatrix}$

$X = \begin{bmatrix}
1 & -\gamma & 0 & 0 \\
0 & 1 & -\gamma & 0 \\
0 & 0 & 1 & -\gamma \\
0 & 0 & 0 & 1 & -\gamma
\end{bmatrix}$

2) What is the 1st selected feature of $X$ (selected for $R$)?

$\begin{bmatrix}
-\gamma & -\gamma^2 & -\gamma^3 \\
-\gamma & -\gamma^2 & -\gamma^3 \\
-\gamma & -\gamma^2 & -\gamma^3 \\
-\gamma & -\gamma^2 & -\gamma^3
\end{bmatrix} = \begin{bmatrix}
-\gamma & -\gamma^2 & -\gamma^3 \\
-\gamma & -\gamma^2 & -\gamma^3 \\
-\gamma & -\gamma^2 & -\gamma^3 \\
-\gamma & -\gamma^2 & -\gamma^3 + 1
\end{bmatrix}$

and if you go through rest of steps works out

Theorem: if $\mathbf{V}^*$ is $m$-sparse in $\bar{\mathbf{z}}$ and $\max_{\mathbf{r}} \| \mathbf{X}^\dagger \mathbf{X}^\dagger \mathbf{X} \|_1 < 1$ (remember $\mathbf{X}^\dagger$ is pseudo-inverse)

for $X = \frac{\bar{\mathbf{z}}}{-\gamma \bar{\mathbf{p}}}$

then OMP-BRM with $\mathbf{B}=0$ will return $\mathbf{w}$ s.t. $\mathbf{V}^* = \mathbf{w}$

In at most $m$ iterations

Note: may be hard to know in advance

Still assuming

$\infty$ data

no noise (but expect can be extended)