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 $\hat{w}$
  $I = \emptyset$
  $w = 0$ (0 vector, length $k$)
  repeat
    $c_i = |X^T(y - Xw)|$
    $j = \arg \max_{i \in I} c_i$
    if $c_j > \beta$, add $j$ to $I$
    $w_I = X^+_I y$
  until $c_j \leq \beta$ or $I = \{j\}$ (all features added)

  $X^+_I \equiv (X^T X)^{-1} X^T$

  comp cost:
  at least $nk^2$ per iteration
  $O(k^3)$ for inverse

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

  thm (Tropp 2004)
  if $y$ is $m$-sparse in $X$ and
  \[ \max_{i \notin opt} \| X_{opt}^T x_i \|_1 \leq 1 \]
  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 oo 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

$V^*$

$\pi V$

$V$

$\pi^*$

\[ \min_{w} \| R + \gamma \pi' w - \pi w \|^2 = \| R - \left( \pi - \gamma \pi \pi' \right) w \|^2 \]

use OMP algorithm where

$y = R$

$X = \pi - \gamma \pi'$

$\text{OMP-TD (OMP w/LSTD)}$

interleave fixed pt calculation given current selected features

with adding features that correlate most with Bellman error

\text{OMP-TD algorithm}

\text{input}

$\pi \in \mathbb{R}^{n \times k}$: $\pi_{i,j} = \gamma_j(s_i)$

$\pi' \in \mathbb{R}^{n \times k}$: $\pi'_{i,j} = \gamma_j(s_i')$

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

$\gamma \in [0,1)$, $\beta \in \mathbb{R}$

\text{output: $w$}

$I \subseteq \{1\}

w = 0

\text{repeat}

$c = \frac{1}{n} \pi^T (R + \gamma \pi' - \pi w - \pi w) \|/n$ //find features correl w/Bellman error

$j = \arg\max_i x \in C$

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

$w_x = \left( \pi - \gamma \pi \pi' \right) x \pi^T R$ //compute fixed point using selected features

until $c_j \leq \beta$ or $I = \emptyset$

\text{guarantees/properties}

\text{OMP-TD}

good news: if feature j adding has a corresponding $c_j > \text{some threshold}$

then it improves a bound on the distance between $V^*$ and the fixed point
but can be suboptimal
consider
\[
\begin{array}{cccccc}
S_1 & S_2 & S_3 & S_4 & S_5 \\
\circ & \circ & \circ & \circ & \circ
\end{array}
\]

rewards:
\[
\begin{array}{cccc}
-\gamma^2 \gamma^3 & 1 & 1 & 1 & 0 \\
0 & 1 & \gamma & 0 & 0
\end{array}
\]

let \( \mathbf{e} \) be an orthonormal basis defined by indicator functions
\[
\varphi_i(s) = 1_{s=s_i}
\]

\( V^* \) is 3-sparse in \( \mathbf{e} \)

\( \mathbf{e} \) opt only requires features 2, 3 \& 4

**Exercise**

at start \( w = 0 \)

1) what is residual vector? \( (R + \gamma \mathbf{e} \cdot \omega \cdot \mathbf{e} \cdot \omega) \)

2) what feature will OMP-TD select first?

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
-\gamma^2 \gamma^3 \\
1 \\
0 \\
0
\end{bmatrix}
= \begin{bmatrix}
-\gamma^2 \gamma^3 \\
1 \\
0 \\
0
\end{bmatrix} \\
\gamma = \frac{65}{100} , \text{ 1st feature}
\]

Problem: implies even if \( V^* \) is \( m \)-sparse, OMP-TD cannot guarantee recovery in \( m \) iterations

e.g. may include extra features

can modify example to make many unnecessary features added

0 what about OMP-BRM?

Lemma 1: if \( V^* \) is \( m \)-sparse in \( \mathbf{e} \) then \( R \) is \( m \)-sparse in \( (\mathbf{e} - \gamma P \mathbf{e}) \)

proof: recall \( V^* = (I - \gamma P)^{-1} R \)

\[
(I - \gamma P)^{-1} R = \begin{bmatrix}
\mathbf{e}_{\text{opt}} & \omega_{\text{opt}} \\
\omega_{\text{opt}} & \mathbf{e}_{\text{opt}} & \omega_{\text{opt}} \\
\omega_{\text{opt}} & \mathbf{e}_{\text{opt}} & \omega_{\text{opt}} & \omega_{\text{opt}}
\end{bmatrix}
\]

assume other direction, that have found a \( m \)-sparse rep of \( R \) with param \( \omega' \text{opt} \) in basis \( (\mathbf{e}_{\text{opt}} - \gamma P \mathbf{e}_{\text{opt}}) \)

then

\[
R = \begin{bmatrix}
\mathbf{e}_{\text{opt}} & \omega_{\text{opt}} \text{opt} \\
\omega_{\text{opt}} & \mathbf{e}_{\text{opt}} & \omega_{\text{opt}} \text{opt} \\
\omega_{\text{opt}} & \mathbf{e}_{\text{opt}} & \omega_{\text{opt}} & \omega_{\text{opt}} \text{opt}
\end{bmatrix}
\]

\[
(I - \gamma P)^{-1} R = \begin{bmatrix}
\mathbf{e}_{\text{opt}} & \omega_{\text{opt}} \text{opt} \\
\omega_{\text{opt}} & \mathbf{e}_{\text{opt}} & \omega_{\text{opt}} \text{opt} & \omega_{\text{opt}} \text{opt}
\end{bmatrix}
\]

\[
V = \begin{bmatrix}
\mathbf{e}_{\text{opt}} & \omega_{\text{opt}} \text{opt} \\
\omega_{\text{opt}} & \mathbf{e}_{\text{opt}} & \omega_{\text{opt}} & \omega_{\text{opt}} \text{opt}
\end{bmatrix}
\]

\[
\Rightarrow \text{if perform OMP on } (\mathbf{e}_{\text{opt}} - \gamma P \mathbf{e}_{\text{opt}}) \text{ basis on target } R, \text{ then indices}
\]

of \( (\mathbf{e}_{\text{opt}} - \gamma P \mathbf{e}_{\text{opt}}) \) skeptical yield indices of \( \mathbf{e}_{\text{opt}} \) that should be selected to rep \( V \) (and \( \omega' \text{opt} \text{opt} \) too)