Greedy Feature Selection for Supervised Learning

- matching pursuit (Mallat & 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

  $I \leftarrow \{\}$

  $w \leftarrow 0$ (0 vector, length k)

  repeat

  $c = |X^T (\overbrace{y - Xw}^{\text{residual}})|$

  $j = \text{argmax}_{i \notin I} c_i$

  if $c_j > \beta$, add j to I

  $w_I = X_I^{\dagger} y$

  until $c_j \leq \beta$ or $I = \{\}$ (all features added)

  defn: $X^{\dagger} \equiv (X^T X)^{-1} X^T$

  comp cost: at least $n \times k$ per iteration; $O(k^3)$ for inverse

- OMP properties

  define y as m-sparse in X if $\exists$ 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 (1)

  $\max_{i \notin opt} \| X_{opt}^{\dagger} x_i \|_1 < 1$ (2)

  then OMP called with X, y and $\beta = 0$ will return w_{opt} in m iterations.

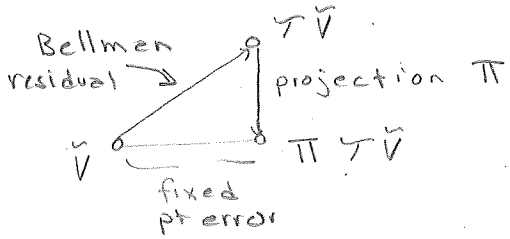
  note: any orthogonal basis X satisfies equation 2.

  the above (I believe) holds for ∞ 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



for now be loose w/samples vs full state space but empirically use samples

- Bellman residual min w/OMP

$$\min_{\omega} \| R + \gamma \Phi' \omega - \Phi \omega \|^2 = \| R - (\Phi - \gamma \Phi') \omega \|^2$$

R ↑ y; $(\Phi - \gamma \Phi')$ = X

use OMP algorithm where

$y = R$

$X = \Phi - \gamma \Phi'$

- 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

$\Phi \in \mathbb{R}^{n \times k}$: $\Phi_{ij} = \varphi_j(s_i)$

$\Phi' \in \mathbb{R}^{n \times k}$: $\Phi'_{ij} = \varphi_j(s_i')$

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

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

output: ω

$I \leftarrow \{\}$

$\omega \leftarrow 0$

repeat

$c = |\Phi^T (R + \gamma \Phi' \omega - \Phi \omega)| / n$ // find features correl w/Bellmen error

$j = \arg\max_{i \notin I} c$

if $c_j > \beta$, $I \leftarrow I \cup \{j\}$

$\omega_I = (\Phi_I^T \Phi_I - \gamma \Phi_I^T \Phi'_I)^{-1} \Phi_I^T R$ // compute fixed point using selected features

until $c_j \leq \beta$ or $\bar{I} = \{\}$

- guarantees/properties
  - OMP-TD
    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

rewards: $-\gamma-\gamma^2-\gamma^3$ 1 1 1 0

V^* 0 $1+\gamma+\gamma^2$ $1+\gamma$ 1 0

let Φ be an orthonormal basis defined by indicator functions

$\varphi_i(s) = \mathbb{1}_{s=s_i}$

V^* is 3-sparse in Φ

Φ_{opt} only requires features 2, 3 & 4

- exercise

at start $\omega = 0$

1) what is residual vector? $(R + \gamma \Phi' \omega - \Phi \omega)$ $R = \begin{bmatrix} -\gamma-\gamma^2-\gamma^3 \\ 1 \\ 1 \\ 0 \end{bmatrix}$

assume have Φ and Φ' for all states

2) what feature will OMP-TD select first?

$\Phi^T R$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -\gamma-\gamma^2-\gamma^3 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -\gamma-\gamma^2-\gamma^3 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

$\gamma = 0.65$, 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

- what about OMP-BRM?

lemma 1: if V^* is m-sparse in Φ then R is (at least) m-sparse in $(\Phi - \gamma P \Phi)$ (transition model)

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

$$(I - \gamma P)^{-1} R = \Phi_{opt} \omega_{opt}$$

$$R = \Phi_{opt} \omega_{opt} - \gamma P \Phi_{opt} \omega_{opt}$$

$$= (\Phi_{opt} - \gamma P \Phi_{opt}) \omega_{opt} \quad \square$$

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

then $R = (\Phi_{opt'} - \gamma P \Phi_{opt'}) \omega_{opt'}$

$$= (I - \gamma P) \Phi_{opt'} \omega_{opt'}$$

$$(I - \gamma P)^{-1} R = \Phi_{opt'} \omega_{opt'}$$

$$V = \Phi_{opt'} \omega_{opt'}$$

$\Rightarrow$ if perform OMP on $(\Phi - \gamma P \Phi)$ basis on target R, then indices of $(\Phi - \gamma P \Phi)$ selected yield indices of Φ that should be selected to rep V (and ωs are the same)

exercise: some example, run OMP-TD

1) what's X?

$X = \Phi - \gamma P \Phi$

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

$X(s_1, :) = [1\ 0\ 0\ 0\ 0] - \gamma [0\ 1\ 0\ 0\ 0]$

$= [1\ -\gamma\ 0\ 0\ 0]$

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

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

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -\gamma & 1 & 0 & 0 & 0 \\ 0 & -\gamma & 1 & 0 & 0 \\ 0 & 0 & -\gamma & 1 & 0 \\ 0 & 0 & 0 & -\gamma & 1-\gamma \end{bmatrix} \begin{bmatrix} -\gamma - \gamma^2 - \gamma^3 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -\gamma - \gamma^2 - \gamma^3 \\ +\gamma^2 + \gamma^3 + \gamma^4 + 1 \\ 1-\gamma \\ 1-\gamma \\ -\gamma \end{bmatrix} \leftarrow$$

and if go through rest of steps works out

- thm: if V^* is m-sparse in Φ and $\max_{i \notin opt} \| X_{opt}^{\dagger} X_i \|_1 < 1$ remember $X^{\dagger}$ is pseudo inverse for $X = \Phi - \gamma P \Phi$ then OMP-BRM with $B=0$ will return w s.t. $V^* = \Phi w$ in at most m iterations

  note: may be hard to know in advance

still assuming

- ∞ data
- no noise (but expect can be extended)