

# 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  $w$
  - $\mathcal{I} \leftarrow \{\}$
  - $w \leftarrow 0$  (0 vector, length  $k$ )
  - repeat
    - $c = |X^T (y - \underbrace{Xw}_{\text{residual}})|$
    - $j = \arg \max_i c_i$
    - if  $c_j > \beta$ , add  $j$  to  $\mathcal{I}$
    - $w_{\mathcal{I}} = X_{\mathcal{I}}^+ y$
  - until  $c_j \leq \beta$  or  $\mathcal{I} = \{1, \dots, k\}$  (all features added)

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

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

## - OMP properties

define  $y$  as  $m$ -sparse in  $X$  if  $\exists X_{\text{opt}}$  composed of  $m$  columns of  $X$  s.t.  $y = X_{\text{opt}} w_{\text{opt}}$   
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)

$$\max_{i \notin \text{opt}} \|X_{\text{opt}}^+ X_i\|_1 < 1 \quad (2)$$

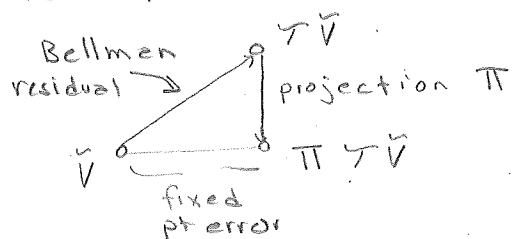
then OMP called with  $X, y$  and  $\beta=0$  will return  $w_{\text{opt}}$   
in  $m$  iterations.

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



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

• Bellman residual min w/OMP

$$\min_w \|R + \gamma \Phi' w - \Phi w\|^2 = \left\| \underset{y}{R} - \underbrace{(\Phi - \gamma \Phi')}_X w \right\|^2$$

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

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

$$\Phi' \in \mathbb{R}^{n \times k} : \Phi'_{ij} = \phi_j(s'_i)$$

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

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

output:  $w$

$$\mathcal{I} \leftarrow \{\}$$

$$w \leftarrow 0$$

repeat

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

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

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

$$w_{\mathcal{I}} = (\Phi_{\mathcal{I}}^T \Phi_{\mathcal{I}} - \gamma \Phi_{\mathcal{I}}^T \Phi'_{\mathcal{I}})^{-1} \Phi_{\mathcal{I}}^T R \quad // \text{compute fixed point using selected features}$$

until  $c_j \leq \beta$  or  $\mathcal{I} = \{1, \dots, k\}$

• guarantees/properties

• 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



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

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

let  $\Phi$  be an orthonormal basis defined by indicator functions

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

$V^*$  is 3-sparse in  $\Phi$

$\Phi_{opt}$  only requires features 2, 3 & 4

o exercise

at start  $w = 0$

1) what is residual vector?  $(R + \gamma \Phi' w - \Phi w)$

assume have  $\Phi$  and  $\Phi'$  for all states

2) what feature will OMP-TD select first?

$$R = \begin{bmatrix} -\gamma - \gamma^2 - \gamma^3 \\ \vdots \\ 0 \end{bmatrix}$$

$\Phi^T R$

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

$\gamma \approx 0.65$ , 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

o what about OMP-BRM?

lemma 1: if  $V^*$  is m-sparse in  $\Phi$  then  $R$  is <sup>at least</sup> m-sparse in  $(\Phi - \gamma P \Phi)$

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

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

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

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

$$\text{then } R = (\Phi_{opt}' - \gamma P \Phi_{opt}') w'_{opt} \\ = (I - \gamma P) \Phi_{opt}' w'_{opt}$$

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

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