Fast Parallel Algorithms for Statistical Subset Selection

Sharon Qian, Yaron Singer

Harvard University

Statistical Subset Selection

Goal: Select k out of n features or samples to maximize objective.

Notation: $\mathbf{y}$: data labels, $\mathbf{X}$: feature space, $\mathbf{w}^{(S)}$: weights s.t. $\text{supp}(\mathbf{w}) \subseteq S$, $\mathbf{\Lambda} = \beta^2$ is an isotropic Gaussian prior and σ^2 as variance

- Feature Selection

$$\ell_{\texttt{reg}}(\mathbf{y}, \mathbf{w}^{(S)}) = \|\mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbf{X}_S\mathbf{w}\|_2^2$$

- Bayesian A-optimality

$$\ell_{\texttt{A-opt}}(\mathbf{y}, \mathbf{w}^{(S)}) = \text{Tr}(\mathbf{\Lambda}^{-1}) - \text{Tr}((\mathbf{\Lambda} + \sigma^{-2}\mathbf{X}_S\mathbf{X}_S^T)^{-1})$$

Previous Algorithms

Main drawback is that previous algorithms are **difficult to optimize** or are **highly sequential**.

- LASSO: to select k features, must tune regularization parameter
- Forward Step-wise Regression: while there is a constant factor approximation, sequential nature requires k iterations [2]

Main Question

Are there fast parallel algorithms for statistical subset selection?

$$\boxed{S^\star = \arg\max_{S \subseteq N : |S| \leq k} f(S) = \arg\max_{S \subseteq N : |S| \leq k} \ell(\mathbf{w}^{(S)})}$$

Adaptive Complexity Model

Definition [1]: an algorithm is r**-adaptive** if it makes r rounds of **parallel** function evaluations.

adaptivity = measure of parallel runtime

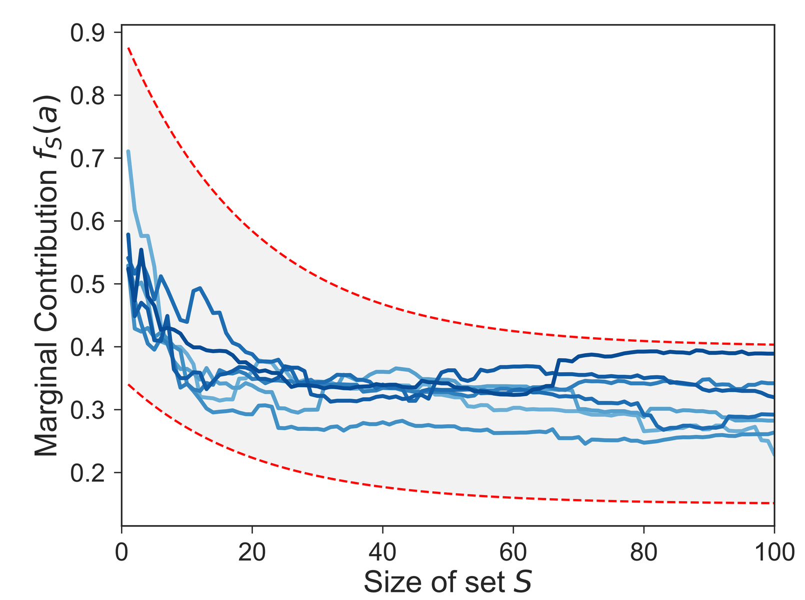
Round 1	Round 2		Round r
$S_{1,1}, f(S_{1,1})$	$S_{2,1}, f(S_{2,1})$		$S_{r,1}, f(S_{r,1})$
$S_{1,2}, f(S_{1,2})$	$S_{2,2}, f(S_{2,2})$	$\cdots$	$S_{r,2}, f(S_{r,2})$
$\vdots$	$\vdots$		$\vdots$
$S_{1,m}, f(S_{1,m})$	$S_{2,m}, f(S_{2,m})$		$S_{r,m}, f(S_{r,m})$

Novel Relaxation of Submodularity

Differential Submodularity

A function $f : 2^N \to_+$ is α-**differentially submodular** for $\alpha \in [0,1]$, if there exist two submodular functions h, g s.t. for any $S, A \subseteq N$, we have that $g_S(A) \geq \alpha \cdot h_S(A)$ and

$$g_S(A) \leq f_S(A) \leq h_S(A)$$



Statistical Subset Selection Objectives are Differentially Submodular

- Feature Selection

$f(S) = \ell_{\texttt{reg}}(\mathbf{w}^{(S)})$ is α-differentially submodular where

$$\alpha = \left(\frac{\lambda_{min}(2k)}{\lambda_{max}(2k)}\right)^2$$

- Bayesian A-optimality

$f(S) = \ell_{\texttt{A-opt}}(\mathbf{w}^{(S)})$ is α-differentially submodular where

$$\alpha = \left(\frac{\beta^2}{\|\mathbf{X}\|^2(\beta^2 + \sigma^{-2}\|\mathbf{X}\|^2)}\right)^2$$

Low Adaptivity Algorithm

Main Theorem

Let f be a monotone, α-**differentially submodular** function where $\alpha \in [0,1]$, then, for any $\epsilon > 0$, DASH is a $\log_{1+\epsilon/2}(n)$ **adaptive algorithm** that obtains the following approximation

$$f(S) \geq (1 - 1/e^{\alpha^2} - \epsilon)\texttt{OPT}.$$

DASH

Uses adaptive sampling technique from submodular maximization.

Input: ground set N, number of iterations r, differential submodularity parameter α

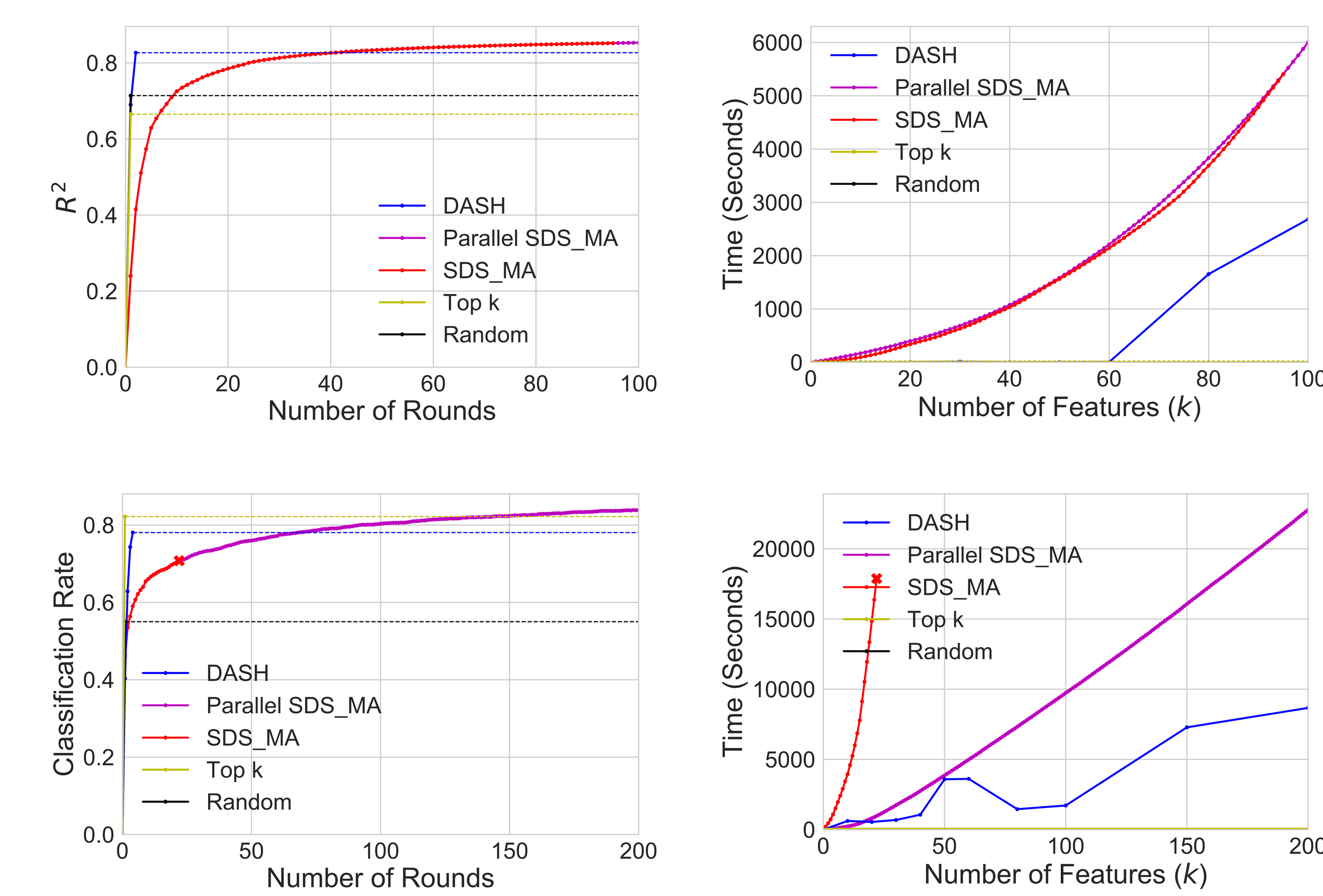
- **Initialize** $S \leftarrow \emptyset$
- **For** r iterations
  - Set threshold $t := (1-\epsilon)(f(O) - f(S))$
  - **While** $\mathbb{E}_{R \sim \mathcal{U}(X)}[f_S(R)] < \alpha^2 \frac{t}{r}$ (sampled set is below threshold)
    - **Discard** elements a from X s.t.

$$\mathbb{E}_{R \sim \mathcal{U}(X)}[f_{S \cup (R \setminus \{a\})}(a)] < \alpha(1 + \frac{\epsilon}{2})t/k\}$$

  - **Add** random sample R to S
- **Return** S

Algorithm in Practice

- DASH achieves comparable solution in fewer rounds compared to traditional methods
- For larger values of k and computationally intensive oracle queries, DASH terminates more quickly



References

[1] Eric Balkanski and Yaron Singer.
The adaptive complexity of maximizing a submodular function.
STOC, 2018.

[2] Ethan R Elenberg, Rajiv Khanna, Alexandros G Dimakis, Sahand Negahban, et al.
Restricted strong convexity implies weak submodularity.
The Annals of Statistics, 46(6B):3539–3568, 2018.