Recursive Sampling for the Nyström Method

Cameron Musco, Christopher Musco ◆ Massachusetts Institute of Technology

Background

The Kernel Method

Adapt any **linear** data analysis method

(regression, principal component analysis, support vector machines, etc.)

to work with **nonlinear** similarity function

How?

These methods only depend on inner product information in the **Gram matrix:**

$$\mathbf{K}_{i,j} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

(n data vectors, $\mathbf{x}_i = \{x_i^1, x_i^2, ..., x_i^d\}$)

Replace Gram matrix and use as usual*

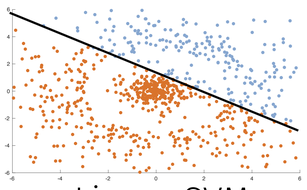
$\mathbf{K}_{i,j} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ → linear inner product

***kernel function needs to be PSD**

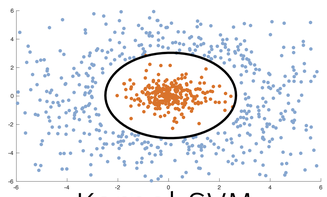
- $\mathbf{K}_{i,j} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ Gaussian kernel inner product
- $\mathbf{K}_{i,j} = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1)^q$ Polynomial kernel inner product
- $\mathbf{K}_{i,j} = \ldots$ Custom nonlinear inner product

The Problem

Pro: Learn nonlinear function classes

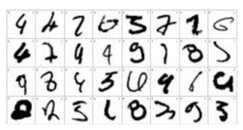


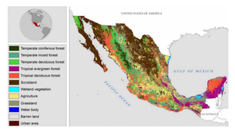
Linear SVM

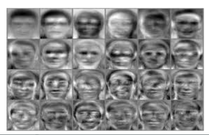


Kernel SVM

Theoretically sound, effective in practice.



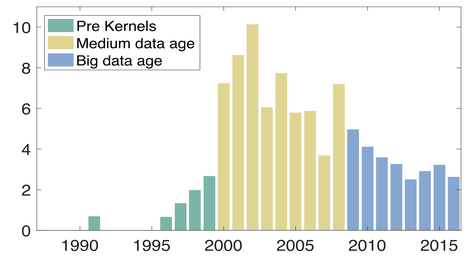




Con: Limited Scalability

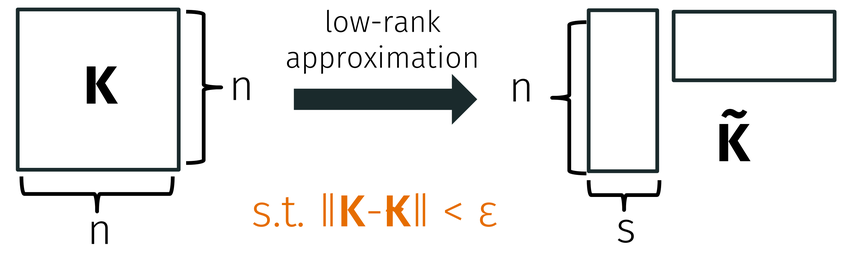
Just writing down **K** takes $O(n^2)$ time!

% of NIPS titles containing "kernel"



Kernel methods can handle high dimensional data, but not large training sets.

Can we use an approximation to K?



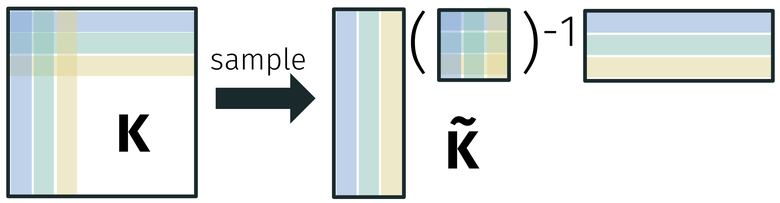
Rank s approximation $\tilde{\mathbf{K}}$ can be stored in O(ns) space, $\tilde{\mathbf{K}}^{-1}$ can be computed in $O(ns^2)$ time, eigendecomposition in $O(ns^2)$ time, etc.

Our Approach

Kernel Approximations	Speed	Accuracy
Incomplete Cholesky / explicit low-rank approximation	$O(n^3)$	High
Random Sketching	$O(n^2)$	High
Random Fourier features	O(n)	Variable
Standard Nyström method	O(n)	Variable
Our Recursive Nyström	O(n)	High

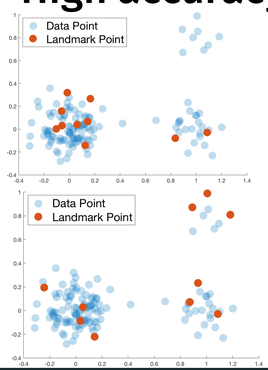
Nyström Method

Low-rank approximation from random sample of the **"landmark"** data points



Time linear in n. Does not require all of **K**!

High accuracy requires better landmarks



Uniformly Random (standard Nyström)

Better Landmarks (importance sampled)

The Nyström method is like "triangulation with noise". For data in s dimensions, we need s landmark points to determine all distances in **K**. If data nearly lies in s dimensions we need O(s) **"well conditioned" points** spread throughout data.

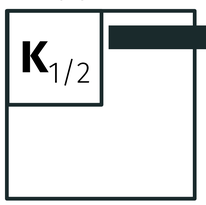
Results

Fast "leverage score" sampling

All good importance sampling probabilities require a good approximation to **K** to compute!

leverage score($\mathbf{x}_i$) = $\mathbf{k}_i^T(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{k}_i$

Main technique: Uniform sampling gives good approximation with many landmarks – e.g. n/2.



Approximate recursively!

log(n) geometrically shrinking levels

$O(ms^2)$ time for level with m points

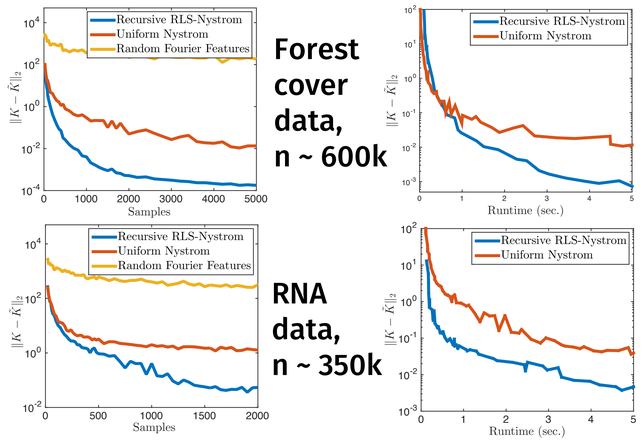
Final runtime = $O(ns^2)$

Strongest theoretical guarantees for approximate kernel learning

# of samples	guarantee
O(statistical dimension)	relative error approx. kernel ridge regression
O(k/ε)	(1+ε) error rank k kernel PCA
O(k/ε)	(1+ε) error kernel k-means clustering

+ general theorems to apply to other kernel problems

State-of-the-art empirical performance



Many more experiments available in the paper.

Simple MATLAB code available at **chrismusco.com**

No tuning required! Works for any kernel.