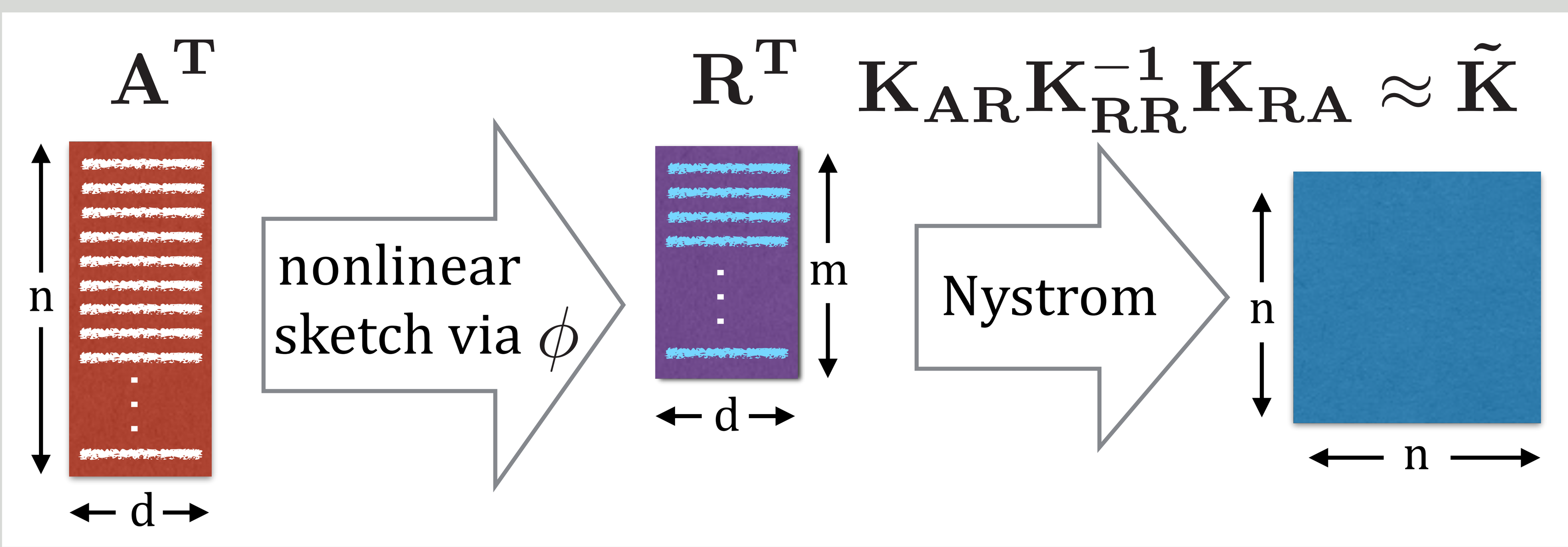
SCHOOL OF COMPUTING
THE UNIVERSITY OF UTAH

Nyström Sketches

Daniel J. Perry[1], Braxton Osting[2], and Ross T. Whitaker[1]
SCI Institute[1] and Department of Mathematics[2], University of Utah, USA
www.sci.utah.edu

Nyström sketching goals

- Desire sketch matrix **R** of size $m \times d$
- **R** should (approximately) minimize projection error in feature space
  - minimize projection error of $\phi(\mathbf{A})$ onto the column span of $\phi(\mathbf{R})$
- **R** found in reasonable time
- Theoretical runtime and error bounds



Objective for Nyström sketching problem

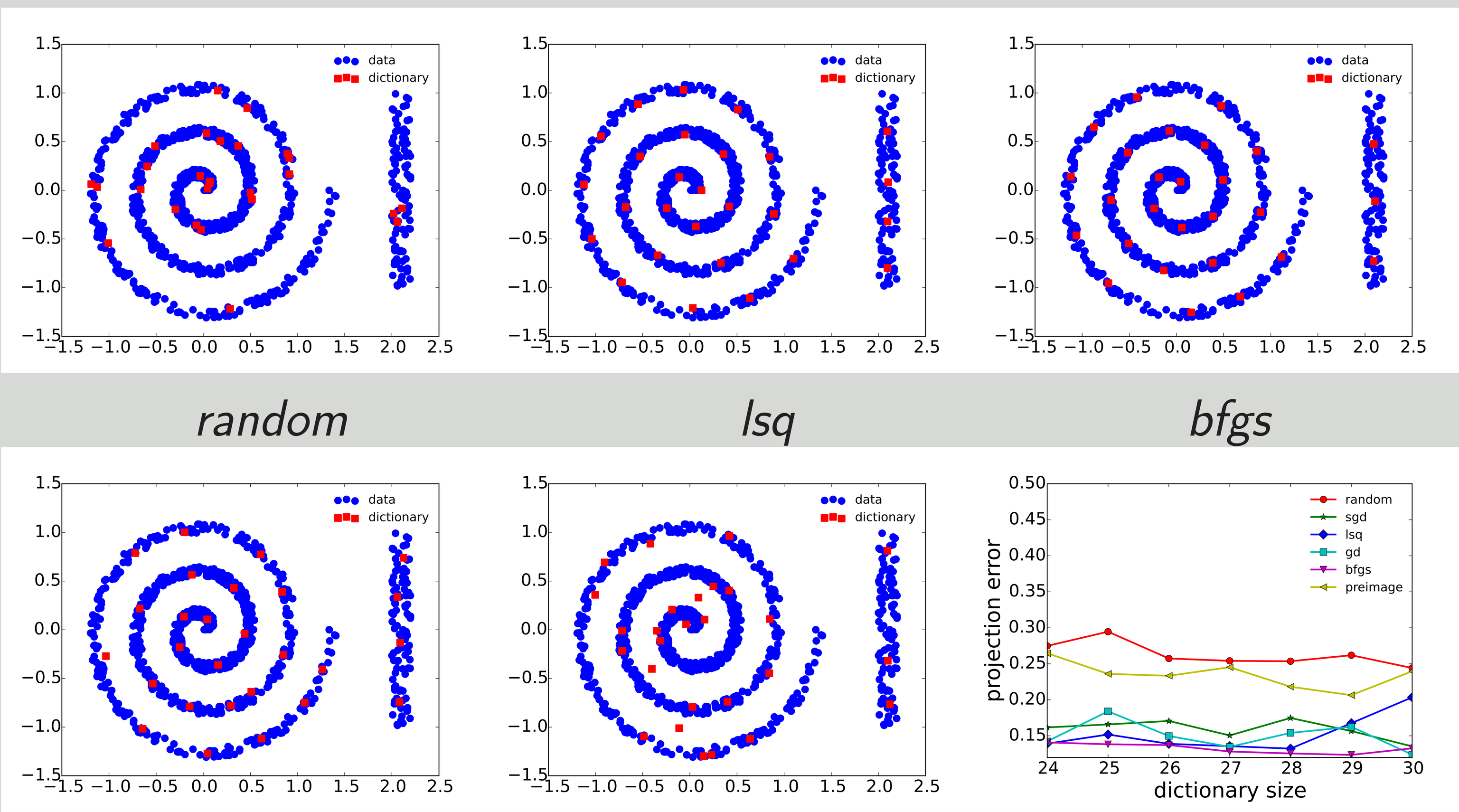
$$\min_{R \in R^{d\times m}} E_{\text{kpca}}(\mathbf{R}) = \frac{1}{n}\sum_{i=1}^{n} \|(\mathbf{I} - \mathbf{V_k V_k^T})\phi(\mathbf{A_i})\|^2, \quad (1)$$

where $\mathbf{V_k}$ is the rank-k kernel PCA basis, $\phi(\mathbf{R}) = \mathbf{V\Sigma U^T}$.

Algorithm: Nyström Sketch

Input Mercer kernel $k(\cdot,\cdot)$, input data A, and sketch size m
Output Nyström sketch $R \in \mathbb{R}^{d\times m}$
Initialize: Let R be a random subset of A
Solve (1) using the Levenberg-Marquardt method (or alternative non-linear least squares method) to find the optimal parameter R.

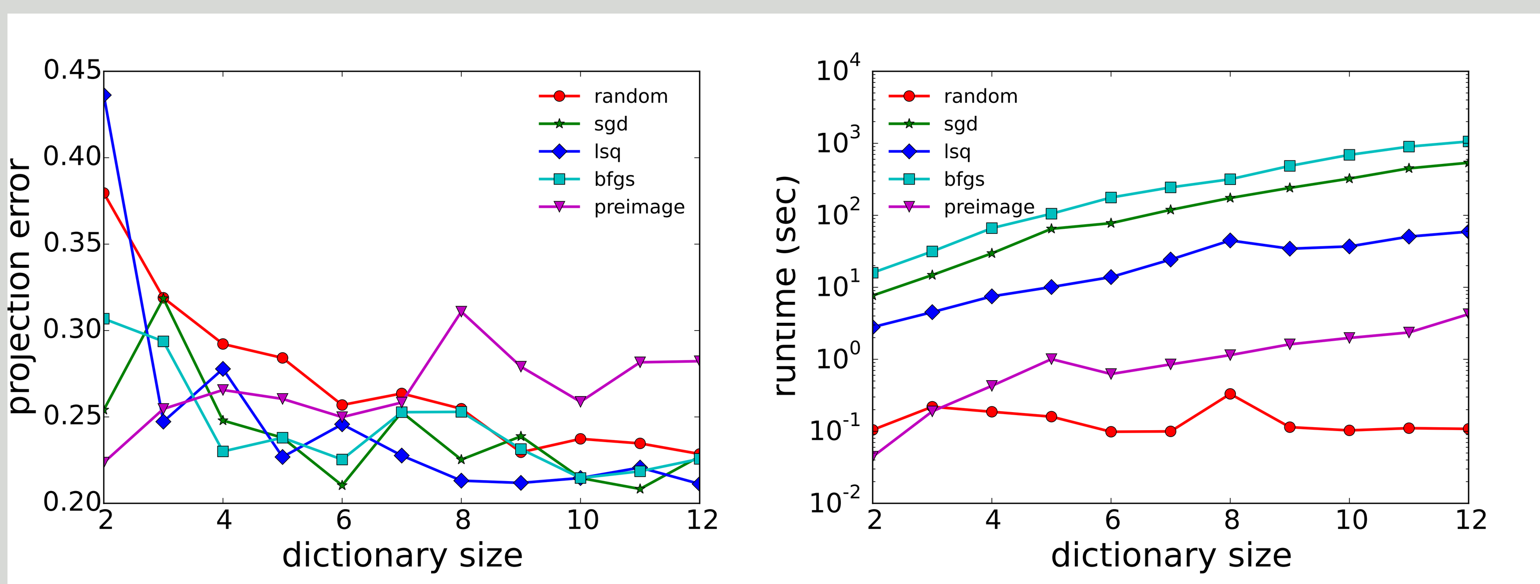
Results: synthetic swirl dataset



random *lsq* *bfgs*

sgd *preimage* comparison

Results: forest dataset



Previous work

- Nyström sketch as a subset
  - *Uniform random sampling* (Williams, et al. 2001), (Drineas, et al. 2005)
  - *Greedy sampling* (Smola, et al. 2002), (Ouimet, et al. 2005)
  - *Optimization* (Alzate, et al. 2008), (Tipping, 2001)
- Nyström sketch as a linear combination
  - *Preimage of kernel PCA basis* (Chin, et al. 2006)

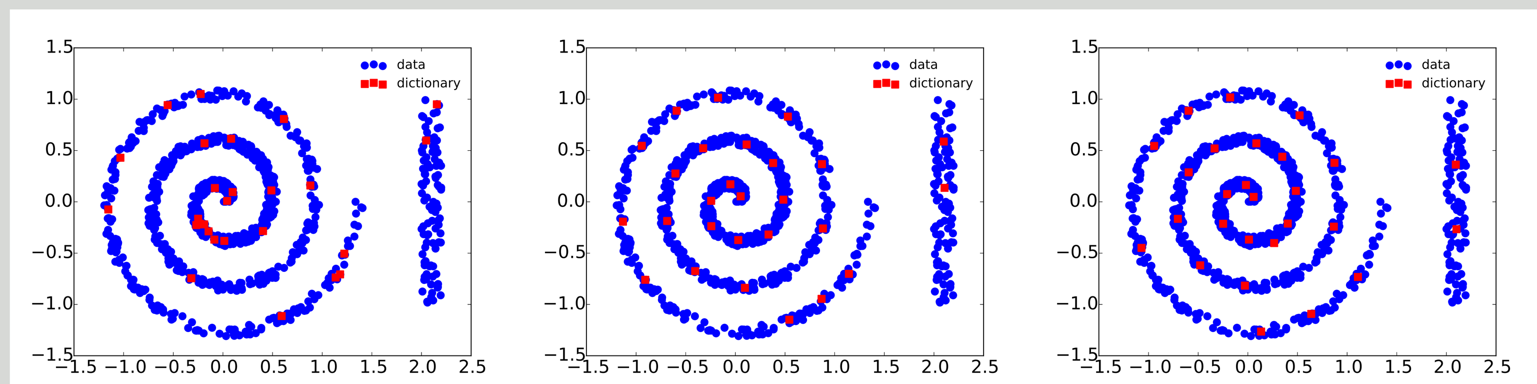
Modifications

- Modify loss function to be *robust* to noise
  -
$$E_{\text{rkpca}}(R) = \frac{1}{n}\sum_{i=1}^{n} \|(I - V_p V_p^T)\phi(A_i)\|_1$$
$$\approx \frac{1}{n}\sum_{i=1}^{n} \sqrt{k(A_i, A_i) - k(R, A_i)^T k(R,R)^{-1} k(R, A_i) + \epsilon} \quad (2)$$
  where the approximation is valid as $\epsilon \to 0$
  - Solve using gradient descent or BFGS instead of a nonlinear least squares
- Modify solution step for use in a *streaming setting*
  - *Solve* (1) *or* (2) *using stochastic gradient descent over the next* b *observed data elements,* $\{A_i, \ldots, A_{i+b}\}$.

Data Sets

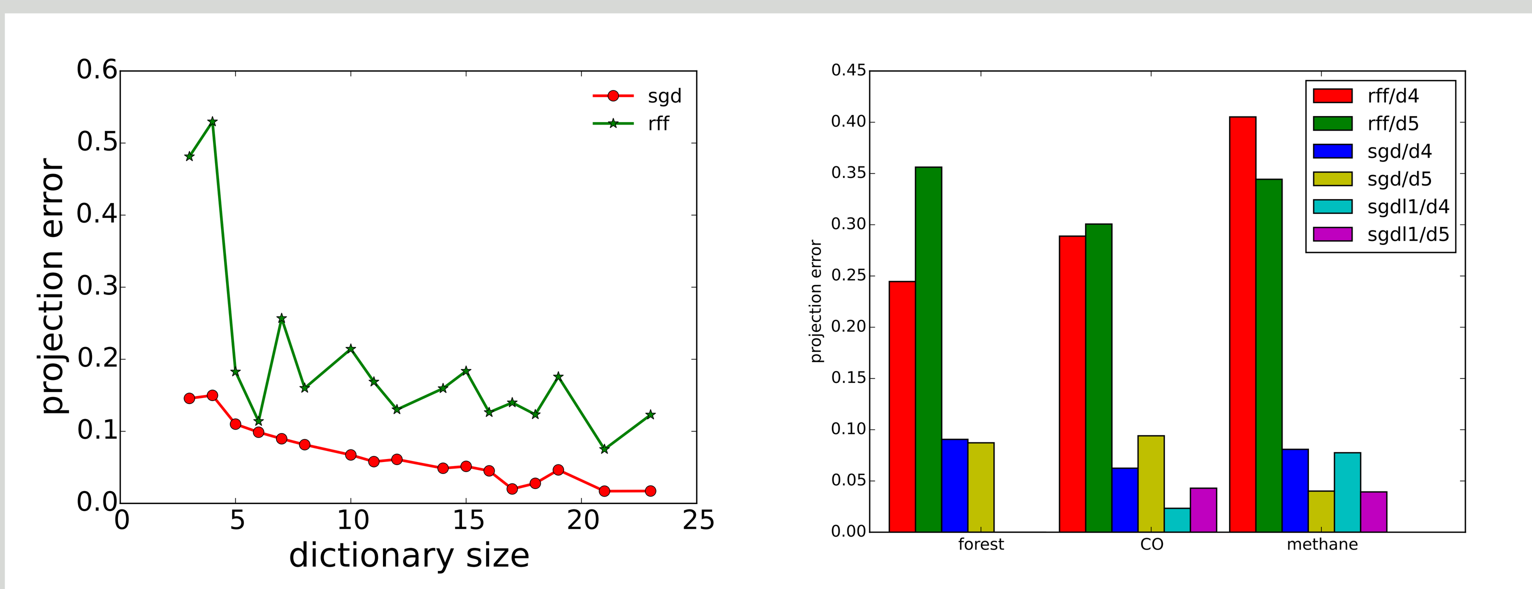
data set	d	n	σ
swirl	2	1000	$\sqrt{0.10}$
forest (subsampled)	54	1000	2.59×10^3
cpu-train	21	6554	5.88×10^5
cpu-test	21	819	5.88×10^5
forest-train	54	522910	2.53×10^3
forest-test	54	58102	2.53×10^3
gas-CO-train	16	3708261	6.44×10^3
gas-CO-test	16	500000	6.44×10^3
gas-methane-train	16	3678504	4.68×10^3
gas-methane-test	16	500000	4.68×10^3

Results: robust formulation on synthetic swirl



sgd *gd* *bfgs*

Results: online formulation



cpu dataset *forest and gas datasets*

http://www.sci.utah.edu {dperry@cs,osting@math,whitaker@cs}.utah.edu