UNIVERSITY OF MINNESOTA
Driven to Discover℠

Scalable Kernel-based Learning via Low-rank Approximation of Lifted Data

Fatemeh Sheikholeslami

Acknowledgement

Prof. Georgios B. Giannakis

NSF grants 1343248, 1442686, 1500713, 1514056, and NIH grant 1R01GM104975-01

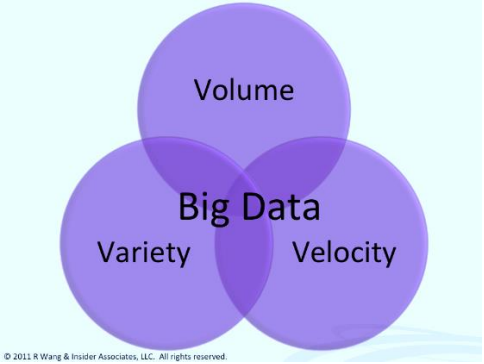
DTC and Dept. of ECE, University of Minnesota

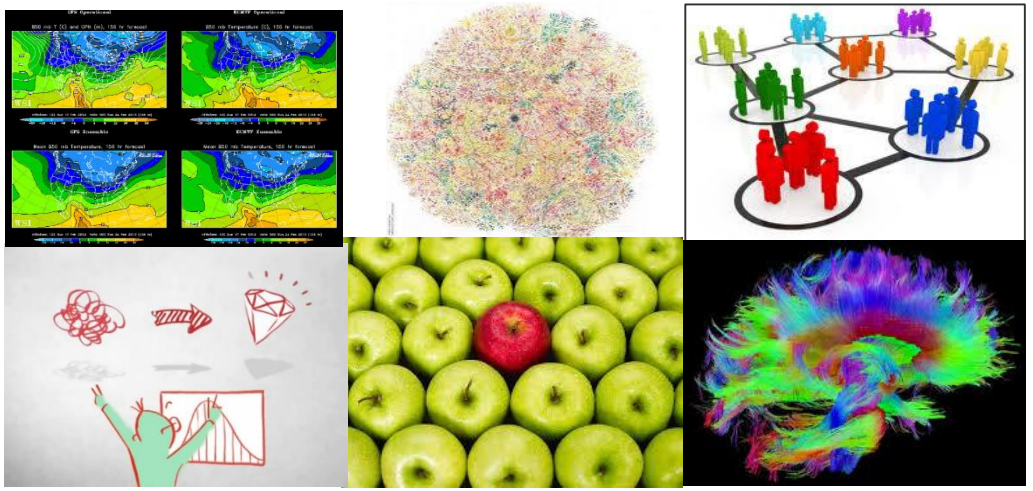
Oct. 4th 2017

55th Allerton Conf. on Comm., Control, and Computing

Learning from “Big Data”

- Challenges
  - Big size ($D \gg$ and/or $N \gg$)
  - Fast streaming
  - Incomplete
  - Noise and outliers
- Opportunities in key tasks
  - Dimensionality reduction
  - Online and robust regression, classification and clustering
  - Denoising and imputation





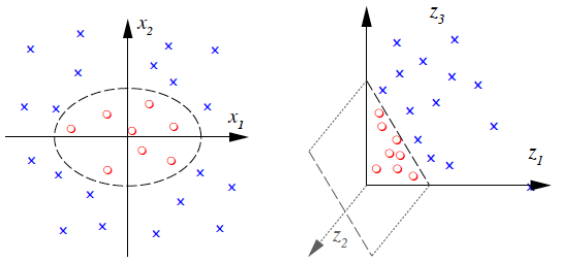
Outline

- Non-linear (kernel-based) learning
- Low rank approximation of lifted data
- Sparsity regularized approximation
- Theoretical guarantees
- Simulation tests

Linear or nonlinear functions for learning?

- **Regression** or **classification**: Given $\{y_n, \mathbf{x}_n\}_{n=1}^N$, find $\hat{f}: \mathbf{x} \rightarrow y = f(\mathbf{x}) + v$
- Lift via nonlinear map $\mathbf{x} \rightarrow \boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^{\bar{D}}$ to linear $y_n = \boldsymbol{\phi}^\top(\mathbf{x}_n)\bar{\boldsymbol{\theta}} + v_n$
  - Pre-select kernel (inner product) function

$$\mathbf{x}_i^\top \mathbf{x}_j \rightarrow k(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}^\top(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j) \text{ e.g., } k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$



$$\Phi: R^2 \rightarrow R^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{(2)}x_1x_2, x_2^2)$$

  - RKHS basis expansion

$$\hat{f}(\mathbf{x}) = \sum_{n=1}^N \alpha_n k(\mathbf{x}, \mathbf{x}_n)$$

- Kernel-based nonparametric ridge regression

$$\mathbf{y} = \mathbf{K}\boldsymbol{\alpha} + \mathbf{v} \rightarrow \boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I}_N)^{-1}\mathbf{y}$$

  - Memory requirement $\mathcal{O}(N^2)$, and complexity $\mathcal{O}(N^3)$

Prior art

❖ Functional gradient descent

- Tractable gradient descent (batch form in primal domain)
  - Pegasus [Shaleve-Schwart et al.'11]; Norma [Kivinen et al.'04]; [Wang et al. '12, Lu et al. '16]

❖ Kernel matrix approximation

- Tractable by leveraging low-rank attribute (in dual domain)
  - Nystrom approximation [Williamson-Seeger'00, Si et al. '14, Wang et al. '14]
  - Incomplete Cholesky decomposition [Fine-Scheinberg'01]

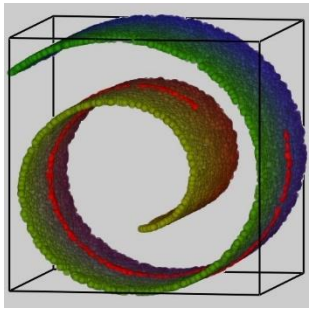
❖ Random feature approximation

- Kernel approximation via feature approximation
  - Random sampling and kernel matrix factorization [Rahimi-Recht'08, Zhang et al.'11]

❖ Reduced-dimensionality features

- Tractability through online extraction of low-dimensional features
  - Online (kernel) PCA [Honnein'012]; linear subspace tracking [Mardani et al.'15]

Low-rank approximation of mapped features



- Recall the nonlinear mapping $\mathbf{x} \to \boldsymbol{\phi}(\mathbf{x})$
- Low-rank approximation $\boldsymbol{\phi}(\mathbf{x}) \simeq \hat{\boldsymbol{\phi}}(\mathbf{x}) := \bar{\mathbf{L}}\mathbf{q}$ where $\bar{\mathbf{L}} \in \mathbb{R}^{\bar{D}\times r}$, $\mathbf{q} \in \mathbb{R}^{r\times 1}$
- Regularized least-squares minimization

$$\min_{\bar{\mathbf{L}},\{\mathbf{q}_i\}_{i=1}^N} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{\phi}(\mathbf{x}_i) - \bar{\mathbf{L}}\mathbf{q}_i\|_{\mathcal{H}}^2 + \frac{\lambda}{N}\sum_{i=1}^{N} \|\mathbf{q}_i\|_2^2$$

- Alternating minimization
  - Update virtual subspace $\bar{\mathbf{L}}$ while fixing feature vectors $\{\mathbf{q}_i\}_{i=1}^N$

$$\Rightarrow \quad \bar{\mathbf{L}} = \boldsymbol{\Phi}_N \mathbf{Q}^\dagger := \boldsymbol{\Phi}_N \mathbf{A} \quad \text{where} \quad \boldsymbol{\Phi}_N := [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)]$$

  - Update vector of extracted features $\{\mathbf{q}_i\}_{i=1}^N$ while fixing subspace $\bar{\mathbf{L}}$

$$\Rightarrow \quad \mathbf{q}_i = (\mathbf{A}^\top \mathbf{K}\mathbf{A} + \lambda \mathbf{I}_r)^{-1}\mathbf{A}^\top \mathbf{k}_i \quad \text{where} \quad \mathbf{K}_{N\times N} := \boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N$$

F. Sheikholeslami and G. B. Giannakis, "Scalable Kernel-based Learning via Low-rank Approximation of Lifted Data," *Proc. of 55th Allerton Conf. on Comm., Control, and Computing*, Oct. 4-6, 2017.

Low-rank appr. with row-sparsity regularization

- Row-sparsity on factor $\mathbf{A} \Rightarrow$ fewer "basis" for the virtual subspace $\bar{\mathbf{L}}$

$$\min_{\mathbf{A},\{\mathbf{q}_i\}_{i=1}^N} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\Phi}_N \mathbf{A}\mathbf{q}_i\|_2^2 + \frac{\lambda}{N}\sum_{i=1}^{N} \|\mathbf{q}_i\|_2^2 + \eta \sum_{i=1}^{N} \|\mathbf{a}_i\|_2$$

- Block coordinate descent
  - Update vector of extracted features $\{\mathbf{q}_i\}_{i=1}^N$ while fixing subspace $\bar{\mathbf{L}}$

    $\Rightarrow \quad \mathbf{q}_i = (\mathbf{A}^\top \mathbf{K}\mathbf{A} + \lambda \mathbf{I}_r)^{-1}\mathbf{A}^\top \mathbf{k}_i$

  - Update rows of subspace factor $\mathbf{A}$ while fixing feature vectors $\{\mathbf{q}_i\}_{i=1}^N$

    $\Rightarrow \quad \mathbf{a}_i[k+1] = \arg\min_{\mathbf{a}_i} \frac{1}{2}\mathbf{a}_i^\top \mathbf{H}_i \mathbf{a}_i + \mathbf{p}_i^\top \mathbf{a}_i + \eta\|\mathbf{a}_i\|_2$ Group shrinkage

Group shrinkage- exact solver

- Target minimization $\mathbf{a}_i[k+1] = \arg\min_{\mathbf{a}_i} F[k] + \eta\|\mathbf{a}_i\|_2$

  where $F[k] := \frac{1}{2}\mathbf{a}_i^\top \mathbf{H}_i \mathbf{a}_i + \mathbf{p}_i^\top \mathbf{a}_i$

- Exact solver

$$\mathbf{a}_i[k+1] = \begin{cases} \mathbf{0} & \text{if } \|\mathbf{p}_i\|_2 \le \eta \\ \left(\mathbf{H}_i + \frac{\eta}{\Delta}\mathbf{I}_r\right)^{-1}\mathbf{p}_i & \text{o.w.} \end{cases}$$

where $\Delta := \|\mathbf{a}_i[k+1]\|$

- Not closed form!
- Root-finding iterations required

Algorithm BCD-Ex

Input: data $\{x_i\}_{i=1}^N$, $I_{\max}$, or, ϵ
Initialize $\mathbf{a}_1 = \mathbf{1}_{r\times 1}$, $\mathbf{a}_i = \mathbf{0}_{r\times 1} \quad \forall i \neq 1$
repeat
  for $i = 1$ **to** N **do**
    $\mathbf{q}_i[k+1] = (\mathbf{A}^\top \mathbf{K}\mathbf{A} + \lambda \mathbf{I}_r)^{-1}\mathbf{A}^\top \mathbf{k}_i$
  end for
  for $i = 1$ **to** N **do**
    Set $\mathbf{H}_i$ and $\mathbf{p}_i$.
    if $\|\mathbf{p}_i\| \le \eta$ **then**
      $\mathbf{a}_i[k+1] = \mathbf{0}$
    else
      $\{\gamma_j, \mathbf{u}_j\}_{j=1}^r = \text{svd}(\mathbf{H}_i)$
      Using root finding algs., find Δ s.t. $\|\mathbf{y}_i(\Delta)\|^2 := \sum_k \frac{(\mathbf{u}_k^\top \mathbf{p}_i)^2}{(\gamma_k \Delta + \eta)^2} = 1$
      $\mathbf{a}_i[k+1] = (\mathbf{H}_i + \frac{\eta}{\Delta}\mathbf{I})^{-1}\mathbf{p}_i$
    end if
  end for
  $k = k + 1$
until $k = I_{\max}$ **or** $\frac{\|\mathbf{Q}[k+1] - \mathbf{Q}[k]\|_F}{\|\mathbf{Q}[k]\|_F} < \epsilon$

Z. Qin, K. Scheinberg, and D. Goldfarb, "Efficient block-coordinate descent algorithms for the group lasso," Mathematical Programming Computation, vol. 5, no. 2, pp. 143–169, 2013.

Group shrinkage- inexact solver

❑ Majorized minimization $\mathbf{a}_i[k+1] = \arg\min_{\mathbf{a}_i} \hat{F}_i(\mathbf{a}_i; \mathbf{a}_i[k]) + \eta\|\mathbf{a}_i\|_2$

$$\hat{F}_i(\mathbf{a}_i; \mathbf{a}_i[k]) := F_i(\mathbf{a}_i[k]) + \langle \mathbf{a}_i - \mathbf{a}_i[k], \nabla F_i(\mathbf{a}_i[k])\rangle + \frac{1}{2\mu_k}\|\mathbf{a}_i - \mathbf{a}_i[k]\|_2^2 \qquad \mu_k \le \|\nabla^2 F_i(\mathbf{a}_i)\|^{-1}$$

➢ Proximal gradient

- Forward step

$$\begin{aligned}\hat{\mathbf{a}}_i[k+1] :=& \mathbf{a}_i[k] - \mu_k \nabla^\top \hat{F}_i(\mathbf{a}_i; \mathbf{a}_i[k])\Big|_{\mathbf{a}_i=\mathbf{a}_i[k]} \\ =& \mathbf{a}_i[k] - \mu_k\Big(\mathbf{H}_i\mathbf{a}_i[k] + \mathbf{p}_i\Big)\end{aligned}$$

- Backward step

$$\mathbf{a}_i[k+1] = \arg\min_{\mathbf{a}_i}\{\frac{1}{2\mu_k}\|\mathbf{a}_i - \hat{\mathbf{a}}_i[k+1]\|_2^2 + \eta\|\mathbf{a}_i\|_2\}$$

$$\mathbf{a}_i[k+1] = \begin{cases}\mathbf{0}, & \text{if } \|\hat{\mathbf{a}}_i[k+1]\| \le \mu_k\eta \\ \Big(1 + \dfrac{\mu_k\eta}{\|\hat{\mathbf{a}}_i[k+1]\| - \mu_k\eta}\Big)^{-1}\hat{\mathbf{a}}_i[k+1], & \text{o.w.}\end{cases}$$

Algorithm BCD-PG

Input: data $\{x_i\}_{i=1}^N$, $I_{\max}$, or, ϵ
Initialize $\mathbf{a}_1 = \mathbf{1}_{r\times 1}$, $\mathbf{a}_i = \mathbf{0}_{r\times 1} \quad \forall i \neq 1$

repeat
  for $i = 1$ **to** N **do**
    $\mathbf{q}_i[k+1] = (\mathbf{A}^\top\mathbf{K}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^\top\mathbf{k}_i$
  end for
  for $i = 1$ **to** N **do**
    set $\mathbf{H}_i$ and $\mathbf{p}_i$.
    $\hat{\mathbf{a}}_i[k+1] = \mathbf{a}_i[k] - \mu_k\Big(\mathbf{H}_i\mathbf{a}_i[k] + \mathbf{p}_i\Big)$
    if $\|\hat{\mathbf{a}}_i[k+1]\| \le \mu_k\eta$ **then**
      $\mathbf{a}_i[k+1] = \mathbf{0}$
    else
      $\Delta = \|\hat{\mathbf{a}}_i[k+1]\| - \mu_k\eta$
      $\mathbf{a}_i[k+1] = (1 + \frac{\eta\mu_k}{\Delta})^{-1}\hat{\mathbf{a}}_i[k+1]$
    end if
  end for
  $k = k+1$
until $k = I_{\max}$ **or** $\frac{\|\mathbf{Q}[k+1] - \mathbf{Q}[k]\|_F}{\|\mathbf{Q}[k]\|_F} < \epsilon$

Convergence and generalization bounds

Proposition 1 Sequences $\{\mathbf{Q}[k], \mathbf{A}[k]\}_{k=1}^{\infty}$ generated by BCD-Ex as well as BCD-PG algorithms converge to a stationary point of the minimization.

- Generalization: performance guarantees on unseen data
  - Consider the equivalent optimization

$$\min_{\mathbf{A}\in\mathcal{A}} \frac{1}{N}\sum_{i=1}^{N} \underbrace{\min_{\mathbf{q}_i\in\mathcal{Q}} \|\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\Phi}_N \mathbf{A}\mathbf{q}_i\|_{\mathcal{H}}^2}_{=:\ \ell(\mathbf{x},\bar{\mathbf{L}})}$$

where $\mathcal{Q} := \{\mathbf{q} \,|\, \mathbf{q}\in\mathbb{R}^r\ , \|\mathbf{q}\|_2 < B\}$

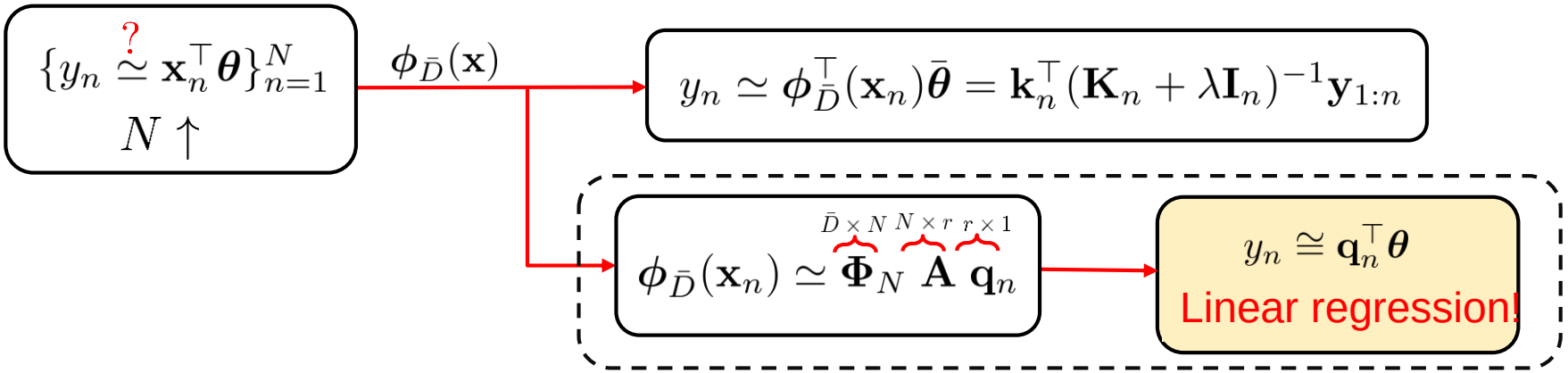
$\mathcal{A} := \{\mathbf{A} | \sum_{i=1}^{N} \|\mathbf{a}_i\|_2 \le \tau\}$

  - Define $\|\mathcal{L}\|_{\mathcal{Q}} := \sup_{\bar{\mathbf{L}}\in\mathcal{L}} \|\bar{\mathbf{L}}\|_{\mathcal{Q}} = \sup_{\bar{\mathbf{L}}\in\mathcal{L}} \sup_{\mathbf{q}\in\mathcal{Q}} \|\bar{\mathbf{L}}\mathbf{q}\|_{\mathcal{H}}$ where $\mathcal{L} := \{\bar{\mathbf{L}} \,|\, \bar{\mathbf{L}} = \boldsymbol{\Phi}_N\mathbf{A}, \mathbf{A}\in\mathcal{A}\}$

Proposition 2. Assume that $\|\mathcal{L}\|_{\mathcal{Q}} \ge 1$, and $\kappa(\mathbf{x},\mathbf{x}) \le \kappa,\ \forall \mathbf{x}$. Then, for a fixed $\delta > 0,$ with probability at least $1-\delta$ we have $\forall \bar{\mathbf{L}} \in \mathcal{L}$

$$\mathbb{E}[\ell(\mathbf{x};\bar{\mathbf{L}})] - 1/N\sum_{i=1}^{N}\ell(\mathbf{x}_i;\bar{\mathbf{L}}) \le \frac{d}{\sqrt{N}}\left(14\kappa B\tau + \frac{\kappa}{2}\sqrt{\ln(16N\kappa^2B^2\tau^2)}\right) + \kappa\sqrt{\frac{\ln(1/\delta)}{2N}}$$

Linearized kernel regression and classification



- Kernel matrix approximation

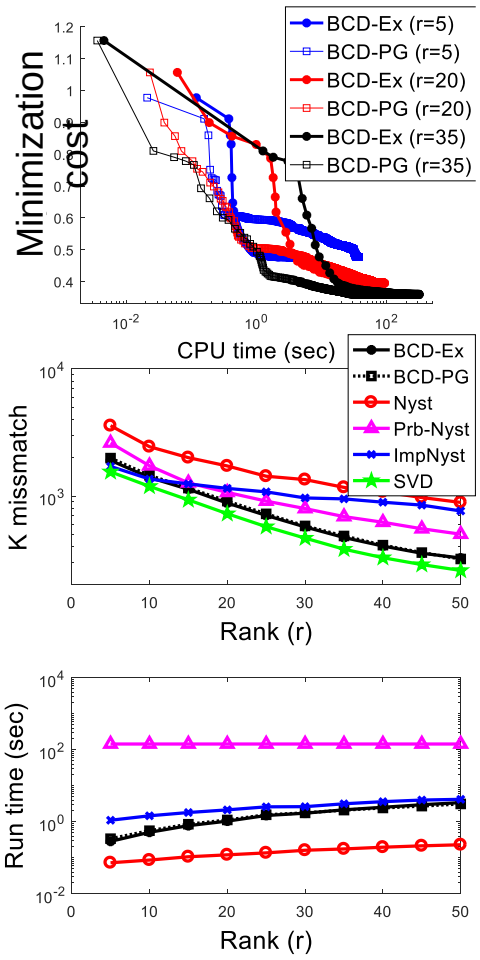
Proposition 3. If $e_i := \|\phi(\mathbf{x}_i) - \hat{\phi}(\mathbf{x}_i)\|_{\mathcal{H}}^2$ iid with $\bar{e} := \mathbb{E}[e_i]$ kernel matrix $\mathbf{K} = \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ can be approximated as $\hat{\mathbf{K}} = \hat{\boldsymbol{\Phi}}^\top \hat{\boldsymbol{\Phi}}$, and w.p. at least $1 - 2e^{-2N\tau^2}$

$$\frac{1}{N}\|\mathbf{K} - \hat{\mathbf{K}}\|_F \leq \sqrt{\bar{e} + \tau}(\sqrt{\bar{e} + \tau} + 2)$$

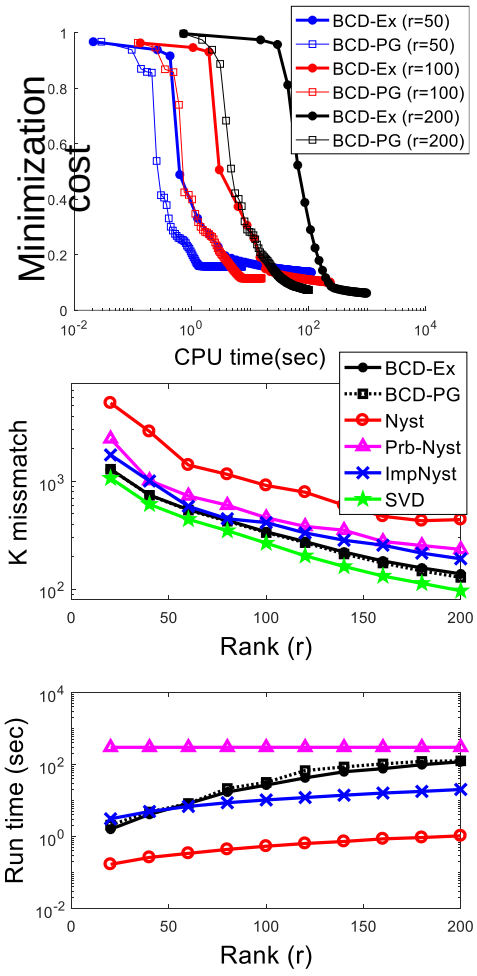
- Bounds also on support vector machines for regression and classification

Test results

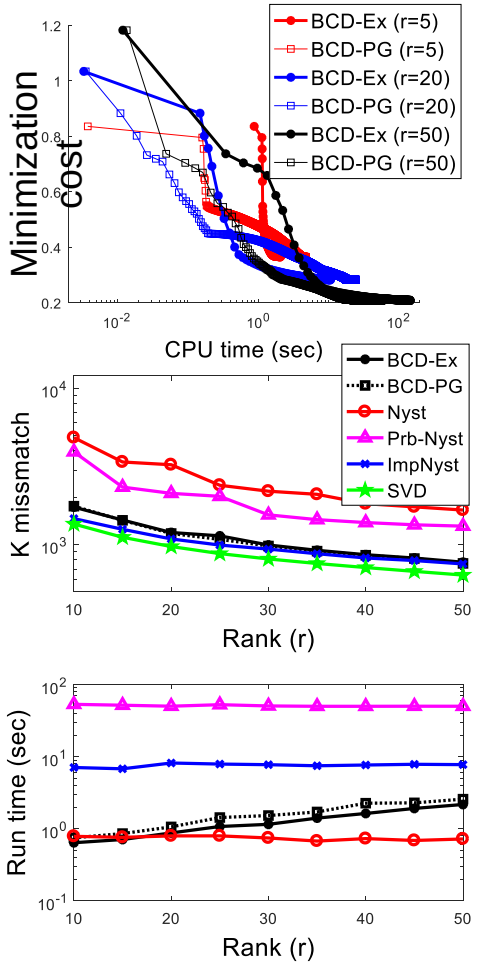
Adult Dataset $N = 32K$



IJCNN Dataset $N = 140K$



Slice Dataset $N = 53K$



Summary

- Kernel-based learning
  - Low-rank approximation of lifted data
  - Sparsity regularized approximation
  - Theoretical guarantees and test results
- Future work
  - Utilization of other regularization techniques
  - Exploitation of censoring to mitigate number of updates
  - Budgeted techniques for bounded memory and complexity growth

Thank you!