Matrix Approximation for Large-Scale Learning

Mehryar Mohri
Courant Institute and Google Research
mohri@cs.nyu.edu

Joint work with
Corinna Cortes, Sanjiv Kumar, Ameet Talwalkar

Motivation

- Kernel-based algorithms:
  - SVMs, Kernel Ridge Regression, KPCA.
  - arbitrary positive definite kernel $K \colon X \times X \to \mathbb{R}$.
  - kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$.
- Computational cost for large-scale problems:
  - $\Omega(n^2)$ space.
  - $O(n^3)$ time for matrix inversion or SVD.

Example

- Invert a large matrix with $n = 18\mathrm{M}$:
  - $\mathbf{K} \approx 1300\mathrm{TB}$.
  - $320{,}000 \times 4\mathrm{GB}$ RAM machines.
- Iterative methods:
  - require matrix-vector products.
  - not suitable for very large dense matrices.
- Sampling-based low-rank approximation:
  - compute and store only $l \ll n$ columns of $\mathbf{K}$.
  - $\longrightarrow$ column-sampling, Nyström method.

This Talk

- Algorithms
- Empirical results
- Guarantees

Low-Rank Approximation

- Positive definite symmetric matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$.
  - singular value decomposition (SVD): $\mathbf{K} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{U}^\top$.
  - best k-rank approximation: $\mathbf{K}_k = \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{U}_k^\top$.
- Objectives
  - find approximation $\widetilde{\mathbf{K}}_k$ of $\mathbf{K}_k$ in linear time with respect to n.
  - minimize reconstruction error $\|\mathbf{K} - \widetilde{\mathbf{K}}_k\|_\xi$, with $\xi = 2, F$.
  - similar loss in learning: $\mathcal{L}_{\mathbf{K}}(h_S) \approx \mathcal{L}_{\widetilde{\mathbf{K}}_k}(h'_S)$.

Nyström Approximation

(Williams & Seeger, 2000; Drineas and Mahoney, 2005)

- Sampling: l columns from $\mathbf{K} \longrightarrow \mathbf{C}$.

$$\begin{array}{cc} & \mathbf{C} \\ \mathbf{K} = & \begin{bmatrix} \mathbf{W} & \mathbf{K}_{21}^\top \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} n \\ & l \end{array}$$

- Approximation: $k \leq l$.

$$\boxed{\widetilde{\mathbf{K}} = \mathbf{C}\mathbf{W}_k^+\mathbf{C}^\top.}$$

when $k = l$, eq. to

$$\mathbf{K}_{22} \rightarrow \mathbf{K}_{21}\mathbf{W}^+\mathbf{K}_{21}^\top.$$

- Computational cost:
  - SVD of $\mathbf{W}$: $O(l^3)$.
  - computation of $\widetilde{\mathbf{K}}$: $O(nlk)$.

Nyström Woodbury Approximation

(Williams & Seeger, 2000)

- Matrix inversion lemma:

$$
\begin{aligned}
&(\lambda\mathbf{I}+\mathbf{K})^{-1}\\
&\approx \big(\lambda\mathbf{I}+\widetilde{\mathbf{K}}\big)^{-1}\\
&= \big(\lambda\mathbf{I}+\mathbf{C}\mathbf{W}_k^{+}\mathbf{C}^\top\big)^{-1}\\
&= \frac{1}{\lambda}\Big(\mathbf{I}-\mathbf{C}\boxed{\big[\lambda\mathbf{I}+\mathbf{W}_k^{+}\mathbf{C}^\top\mathbf{C}\big]}^{-1}\mathbf{W}_k^{+}\mathbf{C}^\top\Big).
\end{aligned}
$$

inversion of an $l \times l$ matrix
instead of an $n \times n$ one.

Applications

- Examples
  - Spectral Clustering (Fowlkes et al., 2004).
  - Kernel Ridge Regression (Cortes, MM, and Talwalkar, AISTATS 2010).
  - Support Vector Machines (Fine and Scheinberg, 2001).
  - Kernel Logistic Regression (Karsmarker et al., 2007).
  - Manifold Learning (Kumar and Talwalkar, 2008).

Large-Scale Manifold Learning

- Nyström Isomap on 18M web faces:
  - largest-scale study to date.
- Visualization suggests good embedding
  - PeopleHopper on Orkut.

Shortest path between images of various celebrities

Fixed Sampling

- Fixed distribution over columns:
  - uniform $O(1)$.
  - diagonal $O(n)$.
  - column-norm $O(n^2)$.
- Empirical results:
  - uniform sampling w/o replacement: best results and fastest for real-world datasets (Kumar, MM, and Talwalkar, 2009).
  - method typically used in practice.

Adaptive Sampling

- Adaptive selection of columns, pre-processing:
  - sparse greedy approximation (Smola and Schoelkopf, 2000).
  - incomplete Cholesky decomposition (Fine and Scheinberg, 2002; Bach and Jordan, 2002).
  - adaptive Nyström (Kumar, MM, and Talwalkar, ICML 2009).
  - k-means (Zhang, Tsang, and Kwok, 2009).
- Results:
  - can achieve better performance.
  - but very costly, no parallelization.

Ensemble Nyström

(Kumar, MM, and Talwalkar, NIPS 2009)

- Sample: $l = pm + s$ columns.
  - p samples $S_1, \ldots, S_p$ of size m.
  - validation sample of size s.
- Approximation: convex combination of p base Nyström approximations.

$$\mathbf{K}_{\text{ens}} = \sum_{k=1}^{p} \mu_k \mathbf{K}_r, \quad \text{with } \boldsymbol{\mu} \in \Delta.$$

- Computational cost: $O(pm^3 + pmkn + C(\boldsymbol{\mu}))$.
  - parallelization: cost similar to single Nyström.

Ensemble Weights

- Uniform: $\mu_r = 1/p$.
- Exponential: $\mu_r = \exp(-\eta \widehat{\epsilon}_r)/Z$, with
  - $\eta \geq 0$ learning parameter, Z normalization factor, and $\widehat{\epsilon}_r$ error of rth expert.
- Regression based weights:

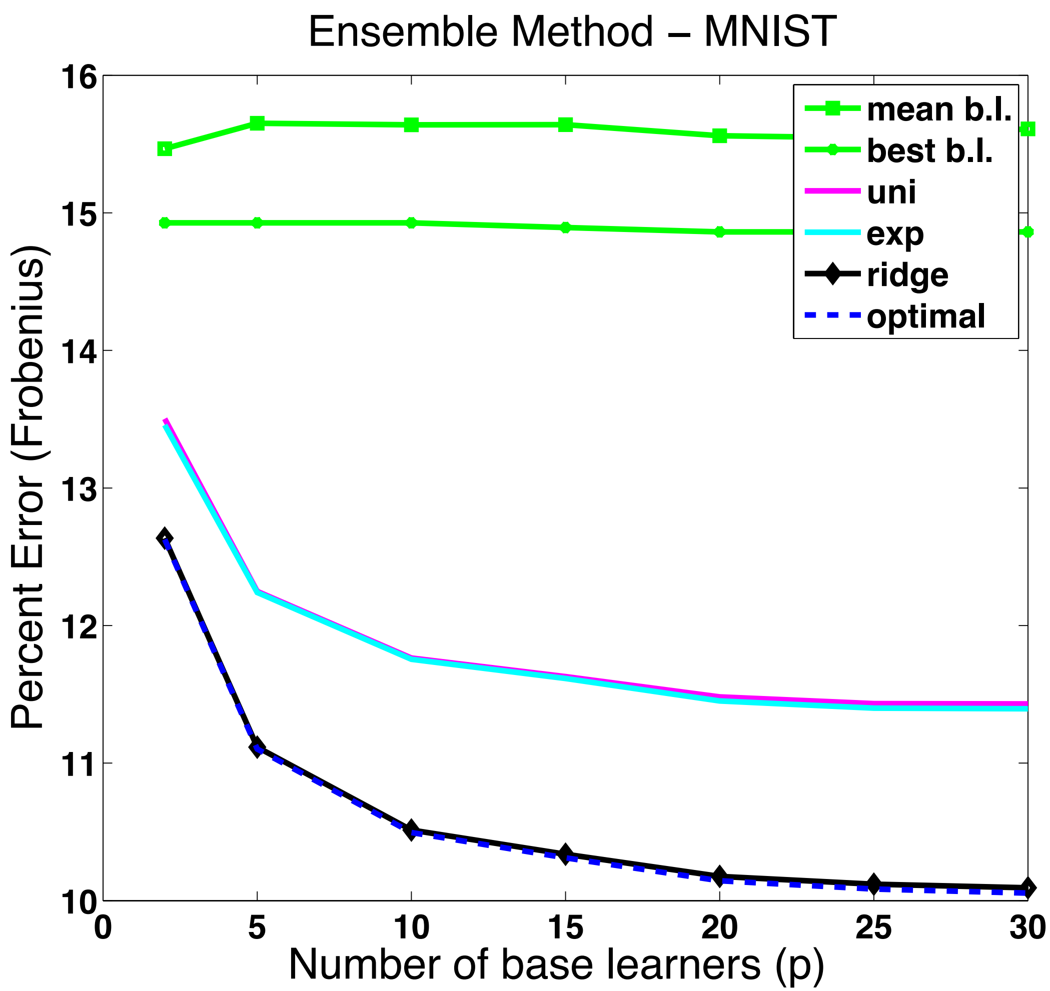
$$\min_{\boldsymbol{\mu} \in \Delta} \ \lambda \|\boldsymbol{\mu}\|_2^2 + \Big\| \sum_{r=1}^{p} \mu_r \widetilde{\mathbf{K}}_r - \mathbf{K}_s \Big\|_F^2 .$$

  - similar with a Lasso-type objective.
  - in practice: non-negativity condition and regularization have small effects.

This Talk

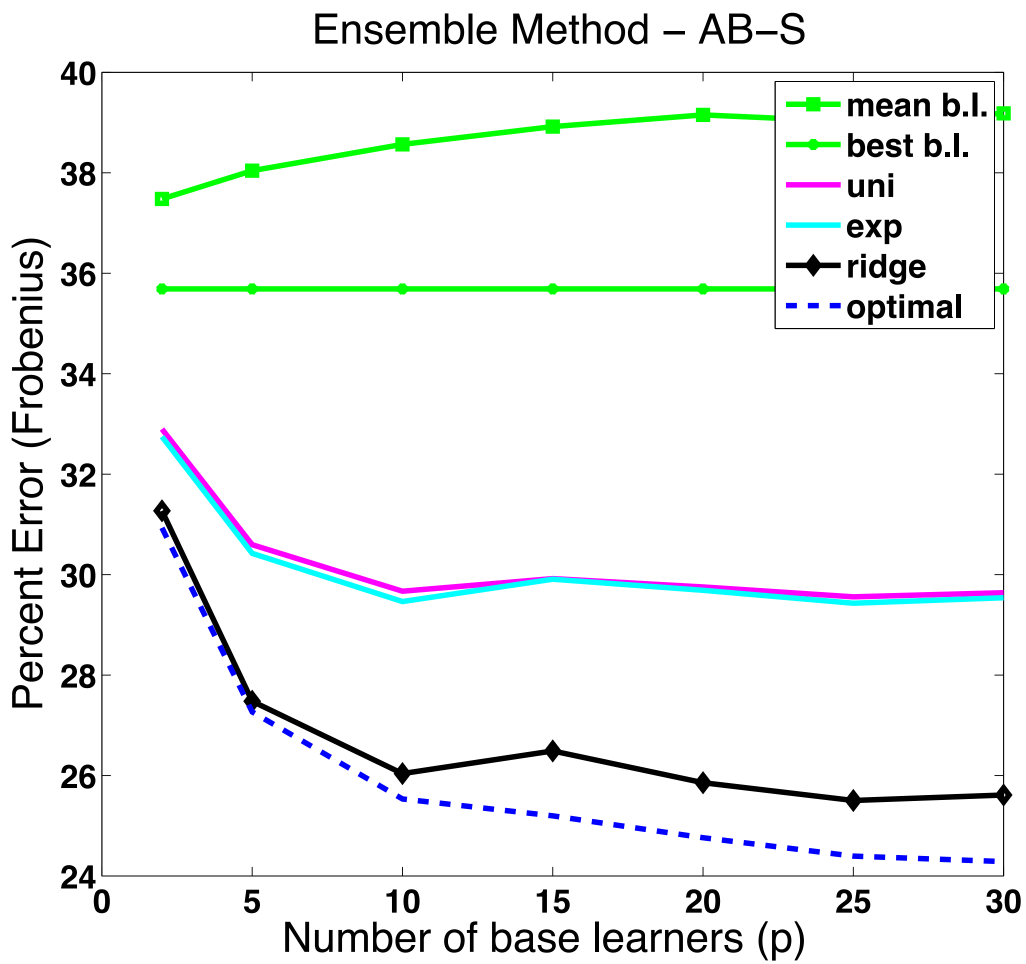
- Algorithms
- Empirical results
- Guarantees

Ensemble Nyström - Experiments

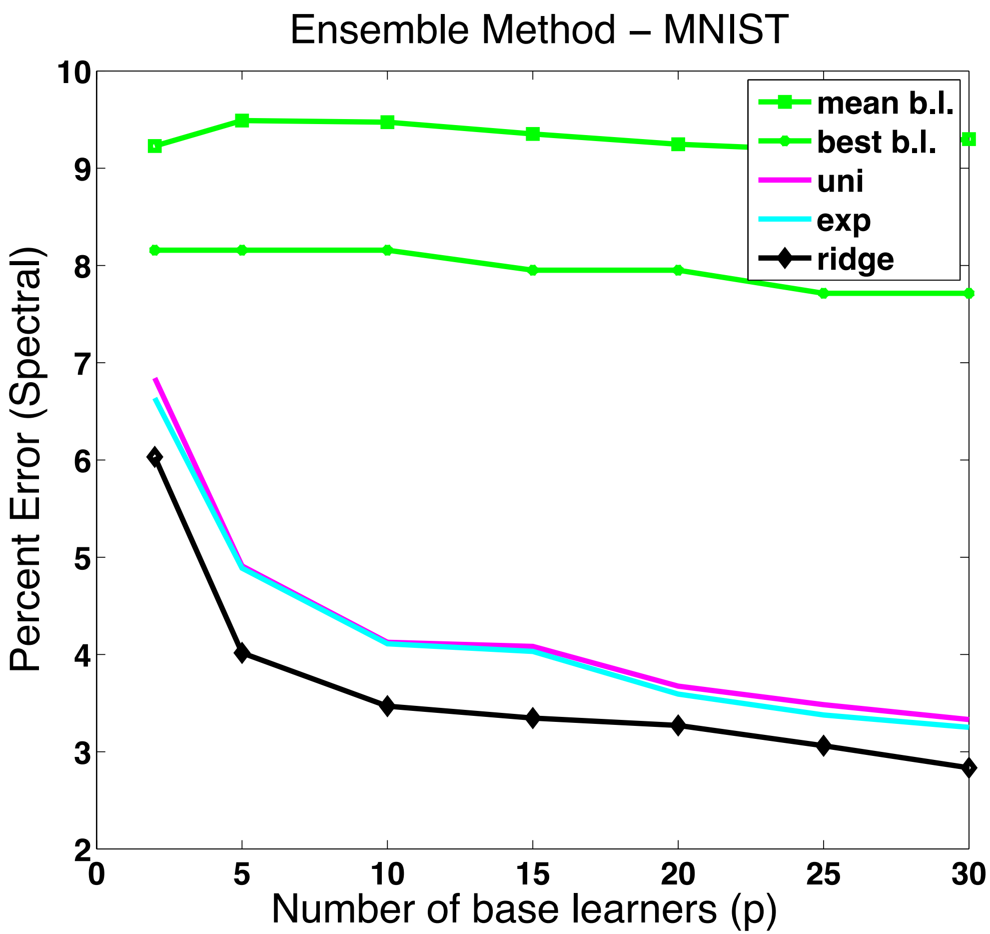


- Relative error: $\dfrac{\|\mathbf{K} - \widetilde{\mathbf{K}}\|_F}{\|\mathbf{K}\|_F}$.

Ensemble Nyström - Experiments

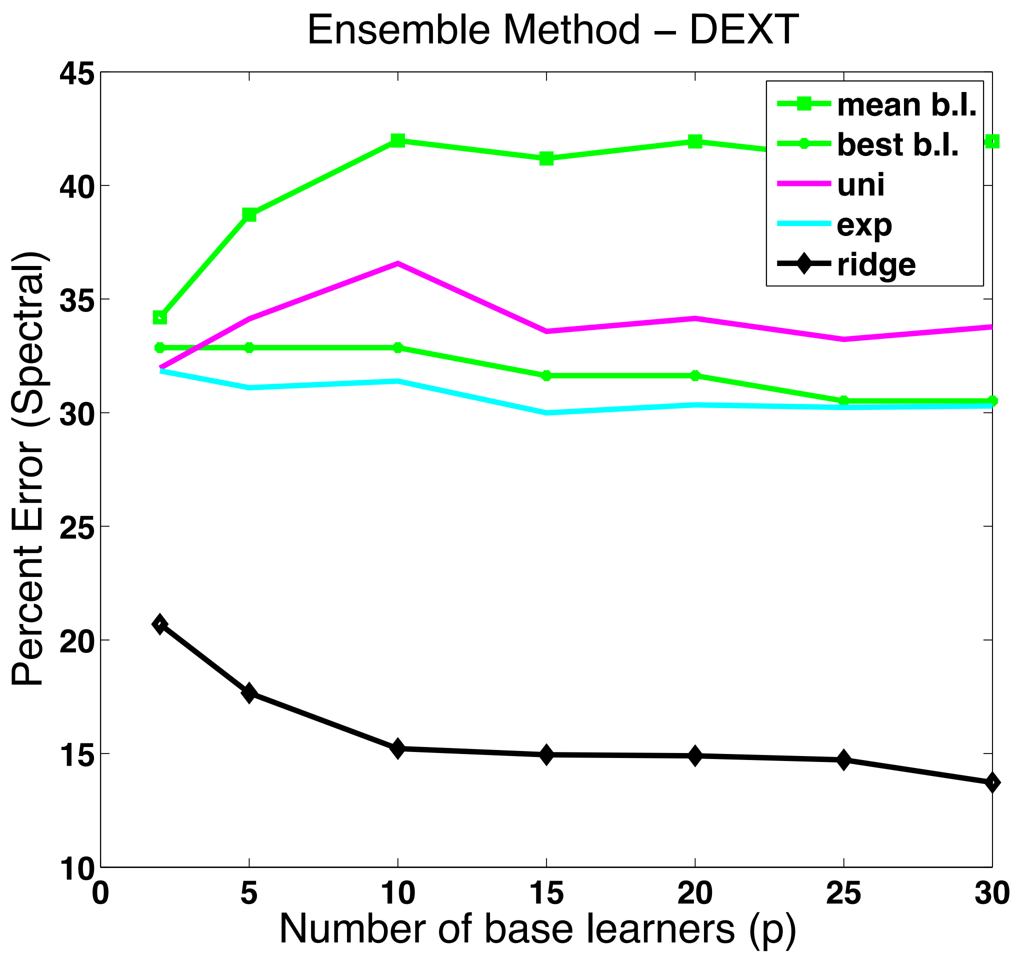


Ensemble Nyström - Experiments

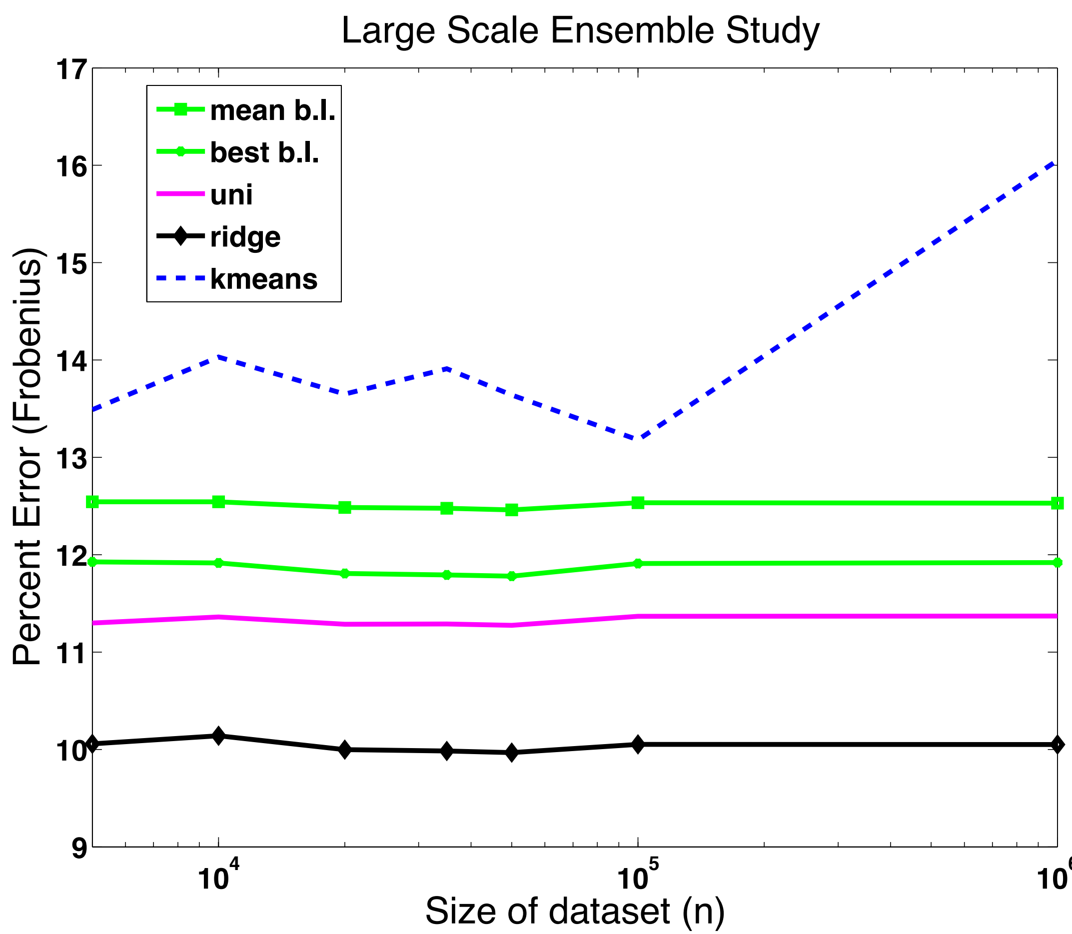


- Relative error: $\dfrac{\|\mathbf{K}-\widetilde{\mathbf{K}}\|_2}{\|\mathbf{K}\|_2}$.

Ensemble Nyström - Experiments



Ensemble Nyström - Experiments



- Fixed-time constraint. 1M points.

This Talk

- Algorithms
- Empirical results
- Guarantees

Nyström Learning Bounds

(Kumar, MM, and Talwalkar, NIPS 2009)

- Theorem: assume that the columns are drawn uniformly w/o replacement. Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\frac{\|\mathbf{K} - \widetilde{\mathbf{K}}_{\text{ens}}\|_2}{\|\mathbf{K}\|_2} \leq \frac{\|\mathbf{K} - \mathbf{K}_k\|_2}{\|\mathbf{K}\|_2} + O\bigg(\frac{1}{\sqrt{m}}\Big(1 + \sqrt{\log \tfrac{1}{\delta}}\Big)\bigg).$$

- Similar bounds for Frobenius norm. More favorable bounds for ensemble Nyström.

Kernel Stability

(Cortes, MM, and Talwalkar, AISTATS 2010)

- Scenario:
  - sample $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$.
  - training with $\mathbf{K}'$ instead of $\mathbf{K}$.
  - testing with the true kernel function.
- Question:
  - how does the use of the approximate kernel $\mathbf{K}'$ affect the learning performance?
  - algorithm-dependent.
  - bounds in terms of $\|\mathbf{K}' - \mathbf{K}\|_\xi$.

Kernel Stability - Ridge Regression

(Cortes, MM, and Talwalkar, AISTATS 2010)

- Optimization problem:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \boldsymbol{\alpha}(\mathbf{K} + \lambda_0 n \mathbf{I})\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top \mathbf{y}.$$

- Theorem: assume that $\max_x (K'(x,x), K(x,x)) \leq R^2$ and $|y| \leq M$, then

$$\forall x \in X, |h'(x) - h(x)| \leq \frac{R^2 M}{\lambda_0^2 n} \|\mathbf{K}' - \mathbf{K}\|_2.$$

- Similar guarantees for several other algorithms: SVMs, SVR, kernel PCA.

Kernel Ridge Regression + Nyström

(Kumar, MM, and Talwalkar, NIPS 2009)

- Theorem: under the same kernel stability assumptions and for uniform sampling w/o replacement, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\forall x \in X, |h'(x) - h(x)| \leq \frac{\kappa M}{\lambda_0^2 m}\Big[\|\mathbf{K} - \mathbf{K}_k\|_2 + \tfrac{m}{\sqrt{n}}\mathbf{K}_{\max}\big(2 + \log \tfrac{1}{\delta}\big)\Big].$$

Conclusion

- Ensemble Nyström algorithm.
  - significant performance improvement.
  - very large-scale experiments.
- Guarantees:
  - Nyström learning bounds.
  - algorithmic kernel stability.
- Better algorithms based on the combination of spectral error and kernel stability.