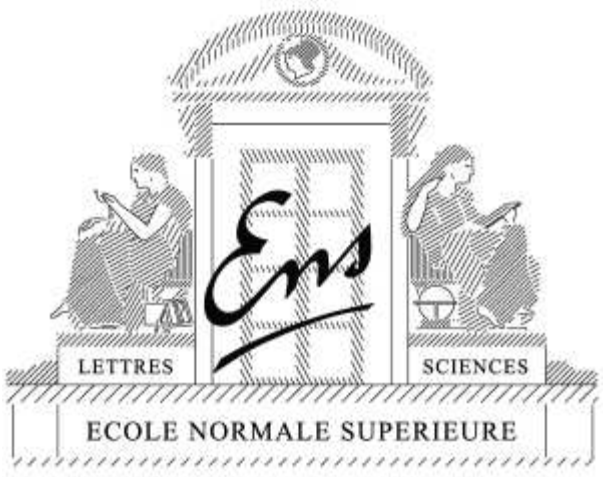
Sharp analysis of low-rank kernel matrix approximations

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France



COLT - June 2013

Why kernels?

- **Provide good abstraction of high-dimensional models**
  - high-dimensional $\approx$ infinite-dimensional
- **Non-linear / non-parametric estimation**
  - Implicitly augment the number of features as n grows
- **Computational complexity**
  - Naive optimization leads to at least $O(n^2)$
- **Goal: lower and upper bounds on complexity**
  - Is it possible to avoid quadratic complexity?
  - Relationships between statistical and computational quantities
  - Beyond worst-case guarantees in $O(1/\sqrt{n})$

Kernel ridge regression

- **Regularized least-squares**: find solution of

$$\min_{f \in \mathcal{F}} \quad \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - \langle f, \Phi(x_i) \rangle \big)^2 \quad + \quad \frac{\lambda}{2} \|f\|^2$$

  - with $\Phi(x_i) \in \mathcal{F}$ Hilbert space

- **Representer theorem**: f is of the form $f = \sum_{i=1}^{n} \alpha_i \Phi(x_i)$

- **Equivalent optimization problem**

  - $K =$ kernel matrix $\in \mathbb{R}^{n \times n}$, $K_{ij} = \langle \Phi(x_i), \Phi(x_i) \rangle = k(x_i, x_j)$

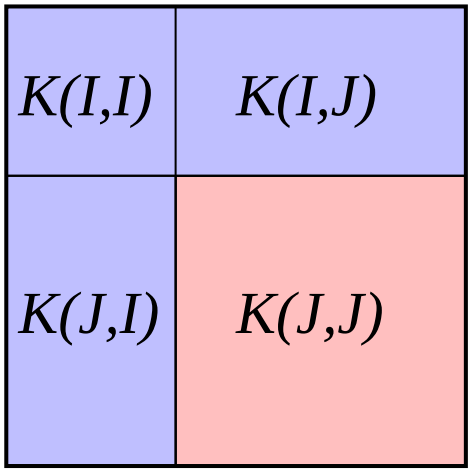
$$\min_{\alpha \in \mathbb{R}^n} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, (K\alpha)_i\big) \quad + \quad \frac{\lambda}{2} \alpha^\top K \alpha$$

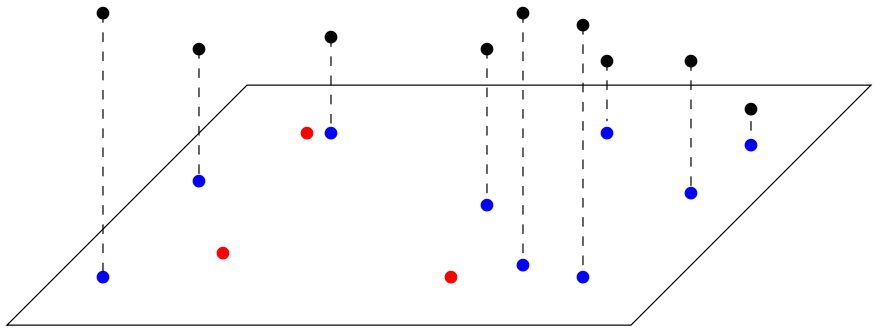
  - $\alpha = (K + n\lambda I)^{-1} y$, predictions $\hat{z} = K(K + n\lambda I)^{-1} y$

Column sampling for kernel matrix approximation

- Given a positive semi-definite matrix $K \in \mathbb{R}^{n \times n}$, and $V = \{1, \ldots, n\}$
  - Approximation for submatrix $K(V, I)$, where $I \subset V$
  - Least-square optimal decomposition:

$$L = K(V,I)K(I,I)^{-1}K(I,V) = k(x_V, x_I)k(x_I, x_I)^{-1}k(x_I, x_V)$$



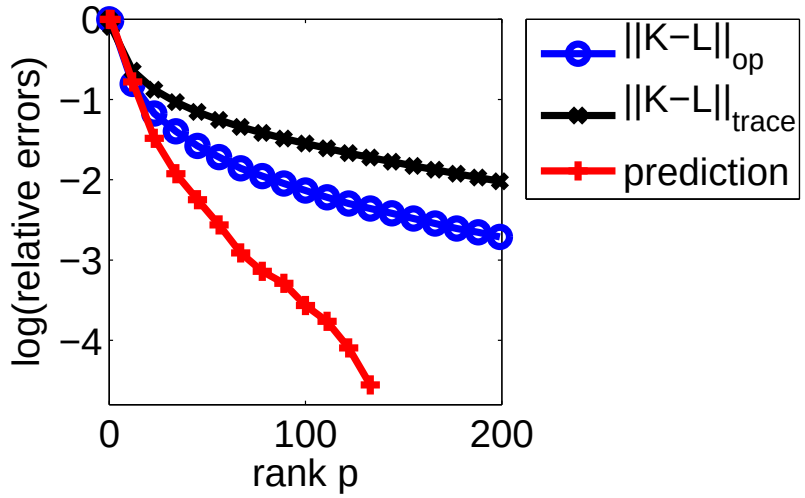


- Computation in $O(|I|^2 n)$ with incomplete Cholesky decomposition
- **Main questions**: size and choice of I (pivoting or random sampling)

Column sampling for kernel matrix approximation

Previous work

- **Bound on $\|K - L\|$**
  - Mahoney and Drineas (2009); Kumar et al. (2012)
  - Tools from matrix concentration inequalities
- **Bound on prediction performance**
  - Non sharp two-step approaches
  - Worst-case performance (Jin et al., 2011)
  - Not taking into account potentially small λ (Cortes et al., 2010)



Fixed design analysis of kernel ridge regression

- $x_1, \ldots, x_n$ deterministic, $y_i = \mathbb{E}y_i + \varepsilon_i = z_i + \varepsilon_i$, $i = 1, \ldots, n$
  - C covariance matrix of ε, prediction $\hat{z} = K(K + n\lambda I)^{-1} y = Hy$
- Bias/variance decomposition of the in-sample prediction error (Wahba, 1990; Hastie and Tibshirani, 1990)

$$\begin{aligned} \tfrac{1}{n}\mathbb{E}_\varepsilon \|\hat{z} - z\|^2 \;=\; & \tfrac{1}{n}\|(H - I)z\|^2 + \;\; \tfrac{1}{n}\operatorname{tr} CH^2 \\ & \operatorname{bias}(K) \quad + \quad \operatorname{variance}(K) \end{aligned}$$

- When $C = \sigma^2 I$, $\operatorname{variance}(K) = \frac{\sigma^2}{n} \operatorname{tr} H^2$.
  - Degrees of freedom $d = \operatorname{tr} H^2$
  - Implicit number of parameters of smoothing matrix H
  - Equal to p, if $\operatorname{rank}(K) = p$ and $\lambda = 0$

Degrees of freedom vs. rank of column sampling approximation

- Column-sampling leads to explicit p-dimensional features
- Degrees of freedom correspond to an implicit number d of parameters
- **What is the link between p and d?**
  - same (or better) performance than full rank problem
- **In general, p must be greater than d**
  - $O(d/n)$ is optimal prediction performance in certain situations
- **Does $p = O(d)$ suffice? Yes!**

Optimal choice of the regularization parameter λ

- **Eigenvalues of** $K = \Theta(n\mu_i)$, $i = 1, \ldots, n$, with $\sum_i \mu_i = \Theta(1)$ so that $\operatorname{tr} K = \Theta(n)$

- **Coordinates of** z on eigenbasis of $K = \Theta(\sqrt{n\nu_i})$ with $\sum_i \nu_i = \Theta(1)$ so that $\frac{1}{n} z^\top z = \Theta(1)$

(μ_i)	(ν_i)	variance	bias	optimal λ	pred. perf.	d	condition
$i^{-2\beta}$	$i^{-2\delta}$	$n^{-1}\lambda^{-1/2\beta}$	λ^2	$n^{-1/(2+1/2\beta)}$	$n^{1/(4\beta+1)-1}$	$n^{1/(4\beta+1)}$	$2\delta > 4\beta+1$
$i^{-2\beta}$	$i^{-2\delta}$	$n^{-1}\lambda^{-1/2\beta}$	$\lambda^{(2\delta-1)/2\beta}$	$n^{-\beta/\delta}$	$n^{1/(2\delta)-1}$	$n^{1/(2\delta)}$	$2\delta < 4\beta+1$
$i^{-2\beta}$	$e^{-\kappa i}$	$n^{-1}\lambda^{-1/2\beta}$	λ^2	$n^{-1/(2+1/2\beta)}$	$n^{1/(4\beta+1)-1}$	$n^{1/(4\beta+1)}$	
$e^{-\rho i}$	$i^{-2\delta}$	$n^{-1}\log\frac{1}{\lambda}$	$(\log\frac{1}{\lambda})^{1-2\delta}$	$\exp(-n^{1/(2\delta)})$	$n^{1/(2\delta)-1}$	$n^{1/(2\delta)}$	
$e^{-\rho i}$	$e^{-\kappa i}$	$n^{-1}\log\frac{1}{\lambda}$	λ^2	$n^{-1/2}$	$\log n/n$	$\log n$	$\kappa > 2\rho$
$e^{-\rho i}$	$e^{-\kappa i}$	$n^{-1}\log\frac{1}{\lambda}$	$\lambda^{\kappa/\rho}$	$n^{-\rho/\kappa}$	$\log n/n$	$\log n$	$\kappa < 2\rho$

References

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

C. Cortes, M. Mohri, and A. Talwalkar. On the impact of kernel approximation on learning accuracy. In *Proc. AISTATS*, 2010.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

R. Jin, T. Yang, M. Mahdavi, Y.-F. Li, and Z.-H. Zhou. Improved bound for the Nyström's method and its application to kernel classification. Technical Report 1111.2262v2, arXiv, 2011.

S. Kumar, M. Mohri, and A. Talwalkar. Sampling methods for the Nyström method. *JMLR*, 13: 981–1006, 2012.

M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.