Sharp analysis of low-rank kernel matrix approximations

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

NIPS Optimization workshop - December 2012

Don’t forget kernels methods!

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

NIPS Optimization workshop - December 2012

Don’t forget kernels methods!
Don’t forget asymptotic analysis!

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

NIPS Optimization workshop - December 2012

Supervised machine learning with convex optimization

Linear vs. non-linear

Small scale vs. large scale

- **1990's - early 2000's**
  - Non-linear kernel methods
  - Non-parametric statistics: convergence rates in $O(n^{-\alpha})$
  - Small-scale problems: complexity in $O(n^2)$ (or more)

Supervised machine learning with convex optimization

Linear vs. non-linear

Small scale vs. large scale

- **1990's - early 2000's**
  - Non-linear kernel methods
  - Non-parametric statistics: convergence rates in $O(n^{-\alpha})$
  - Small-scale problems: complexity in $O(n^2)$ (or more)

- **late 2000's - early 2010's**
  - Linear methods with/without sparsity-inducing regularization
  - Parametric statistics: convergence rates in $O(n^{-1})$ or $O(n^{-1/2})$
  - Large-scale problems: complexity in $O(n)$

Supervised machine learning with convex optimization

Linear vs. non-linear

Small scale vs. large scale

- **1990's - early 2000's**
  - Non-linear kernel methods
  - Non-parametric statistics: convergence rates in $O(n^{-\alpha})$
  - Small-scale problems: complexity in $O(n^2)$ (or more)

- **late 2000's - early 2010's**
  - Linear methods with/without sparsity-inducing regularization
  - Parametric statistics: convergence rates in $O(n^{-1})$ or $O(n^{-1/2})$
  - Large-scale problems: complexity in $O(n)$

- **From naive optimization to naive statistical models**

Outline

- **Introduction**
  - Supervised machine learning and convex optimization
  - Critical review of worst-case analysis
  - Efficient optimization with kernels
- **Classical analysis of kernel ridge regression**
  - Bias / variance
  - Degrees of freedom
- **Sharp analysis of low-rank approximation for kernel methods**
  - Column sampling
  - No loss in predictive performance
- **Choice of regularization parameter**

Supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**

- Prediction $\hat{y} = f(x) = \langle f, \Phi(x) \rangle$, $f \in \mathcal{F} =$ Hilbert space

- **Regularized empirical risk minimization**: find $\hat{f}$ solution of

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

convex data fitting term + regularizer

Supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**

- Prediction $\hat{y} = f(x) = \langle f, \Phi(x) \rangle$, $f \in \mathcal{F}$ = Hilbert space

- **Regularized empirical risk minimization**: find $\hat{f}$ solution of

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

convex data fitting term + regularizer

- Empirical risk: $\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$ training cost

- Expected risk: $R(\theta) = \mathbb{E}_{(x,y)} \ell(y, f(x))$ testing cost

- **Two fundamental questions**: (1) computing $\hat{f}$ and (2) analyzing $\hat{f}$

Supervised machine learning
Worst-case analysis

- Results from Sridharan et al. (2008). See also Boucheron and Massart (2011)

- **Assumptions** ($R =$ expected risk, $\hat{R} =$ empirical risk)
  - $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}(f) + \frac{\lambda}{2}\|f\|^2$
  - $\|\Phi(x)\| \leqslant B$ almost surely
  - L-Lipschitz loss, i.e., R and $\hat{R}$ are LB-Lipschitz continuous

- With probability greater than $1 - \delta$,

$$R(\hat{f}) + \frac{\lambda}{2}\|\hat{f}\|^2 - \min_{f \in \mathcal{F}} \Big\{ R(f) + \frac{\lambda}{2}\|f\|^2 \Big\} \leqslant \frac{16L^2B^2(32 + \log\frac{1}{\delta})}{\lambda n}$$

Supervised machine learning
Worst-case analysis

- Results from Sridharan et al. (2008). See also Boucheron and Massart (2011)

- **Assumptions** ($R =$ expected risk, $\hat{R} =$ empirical risk)
  - $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}(f) + \frac{\lambda}{2}\|f\|^2$
  - $\|\Phi(x)\| \leqslant B$ almost surely
  - L-Lipschitz loss, i.e., R and $\hat{R}$ are LB-Lipschitz continuous

- With probability greater than $1 - \delta$,

$$R(\hat{f}) + \frac{\lambda}{2}\|\hat{f}\|^2 - \min_{f \in \mathcal{F}} \left\{ R(f) + \frac{\lambda}{2}\|f\|^2 \right\} \leqslant \frac{16 L^2 B^2 (32 + \log \frac{1}{\delta})}{\lambda n}$$

- **λ should tend to zero with n!**

Supervised machine learning
Worst-case analysis

- General result with squared norm regularization

$$R(\hat{f}) + \frac{\lambda}{2}\|\hat{f}\|^2 - \min_{f\in\mathcal{F}} \{R(f) + \frac{\lambda}{2}\|f\|^2\} \leqslant O(\frac{1}{\lambda n})$$

- Worst-case: $\lambda = O(n^{-1/2})$

$$R(\hat{f}) - \min_{f\in\mathcal{F}} R(f) \leqslant O(\frac{1}{\sqrt{n}})$$

Supervised machine learning
Worst-case analysis

- General result with squared norm regularization

$$R(\hat{f}) + \frac{\lambda}{2}\|\hat{f}\|^2 - \min_{f \in \mathcal{F}} \left\{ R(f) + \frac{\lambda}{2}\|f\|^2 \right\} \leqslant O\left(\frac{1}{\lambda n}\right)$$

- Worst-case: $\lambda = O(n^{-1/2})$

$$R(\hat{f}) - \min_{f \in \mathcal{F}} R(f) \leqslant O\left(\frac{1}{\sqrt{n}}\right)$$

- For finite dimensional feature spaces $\mathcal{F} = \mathbb{R}^p$
  - **Rates achievable with algorithms of complexity O(pn)**
  - Stochastic gradient and variants

Supervised machine learning
Worst-case analysis

- General result with squared norm regularization

$$R(\hat{f}) + \frac{\lambda}{2}\|\hat{f}\|^2 - \min_{f\in\mathcal{F}} \Big\{R(f) + \frac{\lambda}{2}\|f\|^2\Big\} \leqslant O\Big(\frac{1}{\lambda n}\Big)$$

- Worst-case: $\lambda = O(n^{-1/2})$

$$R(\hat{f}) - \min_{f\in\mathcal{F}} R(f) \leqslant O\Big(\frac{1}{\sqrt{n}}\Big)$$

- **Taking into account the correlation structure of features**
  - All eigenvalues of the kernel matrix and the covariance matrix
  - Between $O(n^{-1})$ and $O(n^{-1/2})$

Why kernels?

- **Finite-dimensional linear models**
  - Efficient optimization algorithms for a fixed λ
  - Choice of λ remains unclear
  - Potential underfitting (parametric statistics)

Why kernels?

- **Finite-dimensional linear models**
  - Efficient optimization algorithms for a fixed λ
  - Choice of λ remains unclear
  - Potential underfitting (parametric statistics)
- **Infinite-dimensional linear models**
  - Few efficient optimization algorithms for a fixed λ
  - Choice of λ remains unclear
  - Implicitly adapt the capacity of predictors as n grows (non-parametric statistics)
  - Higher risk of overfitting
- In many situations, high-dimensional models and infinite-dimensional models exhibit same issues

Why kernels?

- **Provides good abstraction of high-dimensional models**
- **Non-linear estimation**
  - Computer vision, bioinformatics, neuro-imaging
  - Implicitly augment the number of features as n grows
- **Computational complexity**
  - Naive optimization above $O(n^2)$

Why kernels?

- **Provides good abstraction of high-dimensional models**
- **Non-linear estimation**
  - Computer vision, bioinformatics, neuro-imaging
  - Implicitly augment the number of features as n grows
- **Computational complexity**
  - Naive optimization above $O(n^2)$
- **Lower and upper bounds on complexity**
  - Is it possible to avoid quadratic complexity with non-parametric kernel methods?
  - Both theoretical and practical issues

Supervised learning with kernels

- **Regularized empirical risk minimization**: find $\hat{f}$ solution of

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

- **Representer theorem** (Kimeldorf and Wahba, 1971): f may be expressed as $\sum_{i=1}^{n} \alpha_i \Phi(x_i) \Rightarrow f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$
  - Positive definite kernel $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$

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

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

Efficient algorithms for kernel machines
Subquadratic running-time complexity - I

- **Forbidden to compute the kernel matrix**

- **Stochastic gradient with cost $O(t)$ at iteration t leads to $O(n^2)$**
  - Hilbert space iteration: $f_t = (1-\lambda\gamma_t) f_{t-1} - \gamma_t \ell'(y_t, f_{t-1}(x_t))\Phi(x_t)$
  - f_t represented as $\sum_{i=1}^t \alpha_t^i \Phi(x_i)$
  - $\alpha_t^t = -\gamma_t \ell'\big(y_t, \sum_{i=1}^{t-1} \alpha_{t-1}^i k(x_i, x_t)\big)$ and $\alpha_t^{1:t-1} = (1-\lambda\gamma_t)\alpha_{t-1}^{1:t-1}$

Efficient algorithms for kernel machines
Subquadratic running-time complexity - I

- **Forbidden to compute the kernel matrix**

- **Stochastic gradient with cost $O(t)$ at iteration t leads to $O(n^2)$**
  - Hilbert space iteration: $f_t = (1-\lambda\gamma_t) f_{t-1} - \gamma_t \ell'(y_t, f_{t-1}(x_t))\Phi(x_t)$
  - f_t represented as $\sum_{i=1}^t \alpha_t^i \Phi(x_i)$
  - $\alpha_t^t = -\gamma_t \ell'\big(y_t, \sum_{i=1}^{t-1} \alpha_{t-1}^i k(x_i, x_t)\big)$ and $\alpha_t^{1:t-1} = (1-\lambda\gamma_t)\alpha_{t-1}^{1:t-1}$

- **Restricted budget of support vectors**
  - Forgetron (Dekel et al., 2005), Projectron (Orabona et al., 2008), BGSD (Wang et al., 2012)
  - Worst-case guarantees

- **Online selection of examples**: LASVM (Bordes et al., 2005)

Efficient algorithms for kernel machines
Subquadratic running-time complexity - II

- **Random features** (Rahimi and Recht, 2007)
  - For kernels of the form $k(x, x') = \mathbb{E}_\omega \big[\Phi_\omega(x)^\top \Phi_\omega(x') \big]$
  - Use explicit features $(\Phi_{\omega_i}(x))_i$ for samples ω_i, $i = 1, \dots, p$
  - Worst-case guarantees

Efficient algorithms for kernel machines
Subquadratic running-time complexity - II

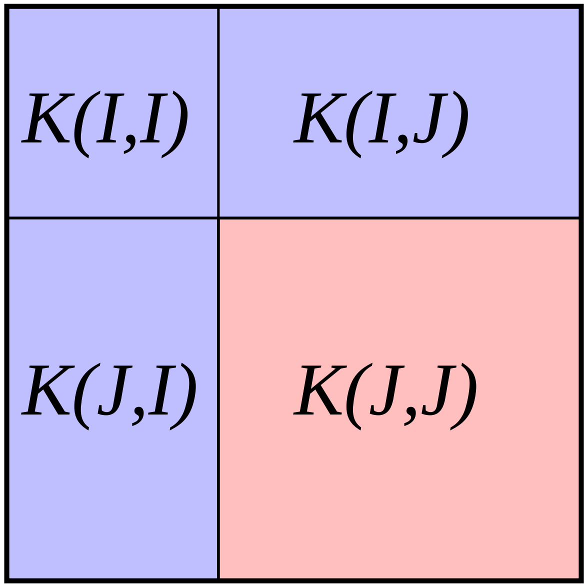
- **Random features** (Rahimi and Recht, 2007)
  - For kernels of the form $k(x,x') = \mathbb{E}_\omega\big[\Phi_\omega(x)^\top \Phi_\omega(x')\big]$
  - Use explicit features $(\Phi_{\omega_i}(x))_i$ for samples ω_i, $i = 1,\dots,p$
  - Worst-case guarantees

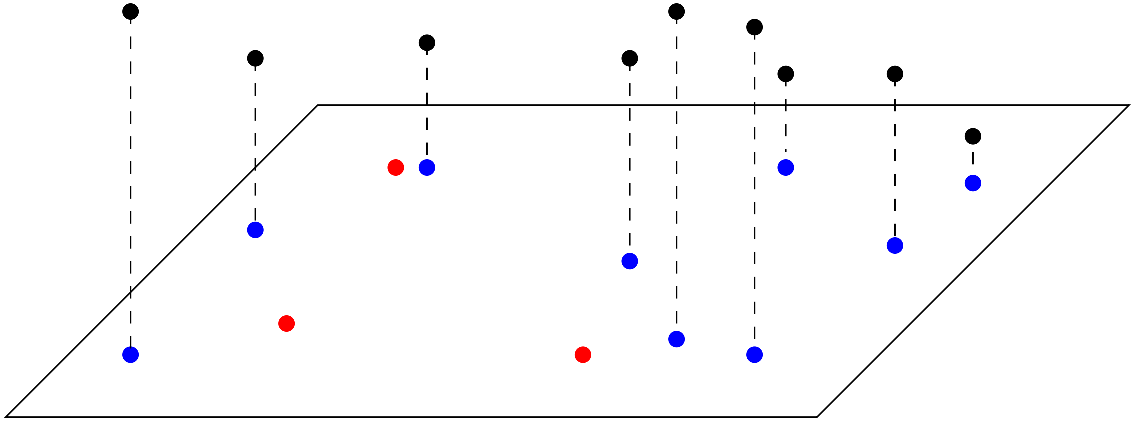
- **Column-sampling**
  - Low-rank approximation of kernel matrix from a subset of its columns/rows
  - Nyström method (Williams and Seeger, 2001), sparse greedy approximations (Smola and Schölkopf, 2000), incomplete Cholesky decomposition (Fine and Scheinberg, 2001), Gram-Schmidt orthonormalization (Shawe-Taylor and Cristianini, 2004), CUR matrix decompositions (Mahoney and Drineas, 2009)

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:

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





- $K(J, J)$ approximated by $K(J,I)K(I,I)^{-1}K(I,J)$

Column sampling for kernel matrix approximation

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

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

  - Property: $K \succcurlyeq L$

- Corresponds to feature map $\tilde{\Phi}(x) = k(x_I, x_I)^{-1/2} k(x_I, x) \in \mathbb{R}^I$

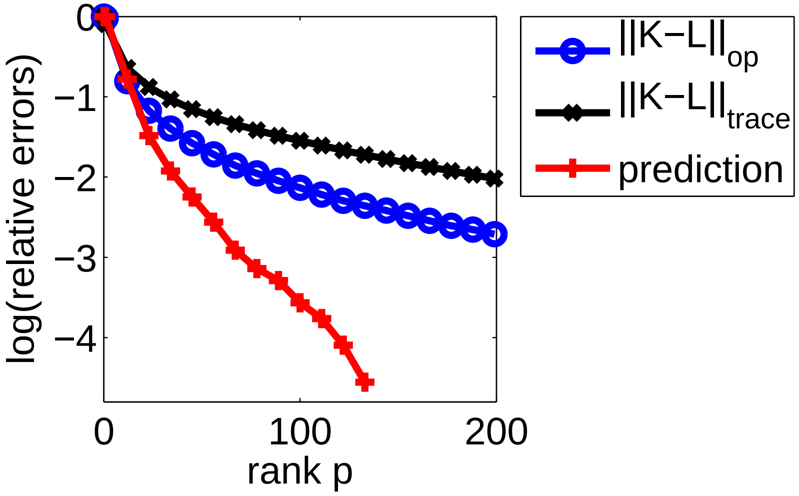
- Computation in $O(|I|^2 n)$ with incomplete Cholesky decomposition

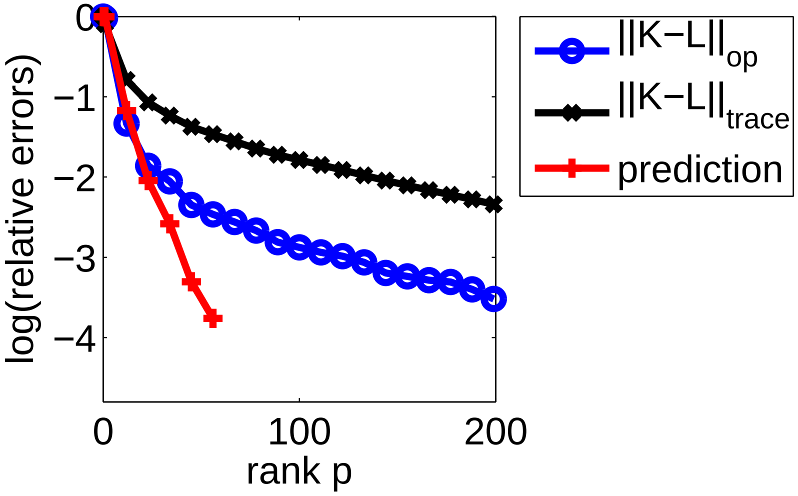
- **Main questions**
  - Choice of I: pivoting or random sampling
  - Cardinality of I

Column sampling for kernel matrix approximation
Previous work

- **Bound on $\|K - L\|$**
  - Mahoney and Drineas (2009); S. Kumar (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)





Outline

- **Introduction**
  - Supervised machine learning and convex optimization
  - Critical review of worst-case analysis
  - Efficient optimization with kernels
- **Classical analysis of kernel ridge regression**
  - Bias / variance
  - Degrees of freedom
- **Sharp analysis of low-rank approximation for kernel methods**
  - Column sampling
  - No loss in predictive performance
- **Choice of regularization parameter**

Kernel ridge regression

- Optimization problem obtained from representer theorem:

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

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2n} \|y - K\alpha\|^2 + \frac{\lambda}{2} \alpha^\top K \alpha$$

- Solution: $\alpha = (K + n\lambda I)^{-1} y$

- Prediction on training data: $K\alpha = K(K + n\lambda I)^{-1} y = Hy$
  - Smoothing matrix H

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; Caponnetto and De Vito, 2007)

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_\varepsilon \|\hat{z} - z\|^2 &= \frac{1}{n}\|\mathbb{E}_\varepsilon \hat{z} - z\|^2 + \frac{1}{n}\operatorname{tr}\operatorname{var}_\varepsilon(\hat{z}) \\
&= \frac{1}{n}\|(H - I)z\|^2 + \frac{1}{n}\operatorname{tr} CH^2
\end{aligned}$$

Fixed design analysis of kernel ridge regression

- $x_1, \dots, x_n$ deterministic, $y_i = \mathbb{E} y_i + \varepsilon_i = z_i + \varepsilon_i$, $i = 1, \dots, n$
  - 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; Caponnetto and De Vito, 2007)

$$\begin{aligned} \frac{1}{n}\mathbb{E}_\varepsilon \|\hat{z} - z\|^2 &= \frac{1}{n}\|\mathbb{E}_\varepsilon \hat{z} - z\|^2 + \frac{1}{n}\operatorname{tr}\operatorname{var}_\varepsilon(\hat{z}) \\ &= \frac{1}{n}\|(H - I)z\|^2 + \frac{1}{n}\operatorname{tr} CH^2 \end{aligned}$$

which may be classically decomposed in two terms:

$$\begin{aligned} \operatorname{bias}(K) &= \frac{1}{n}\|(H - I)z\|^2 = n\lambda^2 z^\top (K + n\lambda I)^{-2} z \\ \operatorname{variance}(K) &= \frac{1}{n}\operatorname{tr} CH^2 = \frac{1}{n}\operatorname{tr} CK^2 (K + n\lambda I)^{-2} \end{aligned}$$

Degrees of freedom

$$\begin{aligned}\text{bias}(K) &= \frac{1}{n}\|(H-I)z\|^2 = n\lambda^2 z^\top (K+n\lambda I)^{-2} z \\ \text{variance}(K) &= \frac{1}{n}\operatorname{tr} CH^2 = \frac{1}{n}\operatorname{tr} CK^2(K+n\lambda I)^{-2}\end{aligned}$$

- When $C = \sigma^2 I$, $\text{variance}(K) = \frac{\sigma^2}{n}\operatorname{tr} H^2 = \frac{\sigma^2}{n}\operatorname{tr} K^2(K+n\lambda I)^{-2}$
- Degrees of freedom: $\operatorname{tr} K^2(K+n\lambda I)^{-2}$ or $\operatorname{tr} K(K+n\lambda I)^{-1}$
  - Implicit number of param. of smoothing mat. $H = K(K+n\lambda I)^{-1}$
  - Equal to p, if $\text{rank}(K) = p$ and $\lambda = 0$

Degrees of freedom

$$\begin{aligned}\text{bias}(K) &= \tfrac{1}{n}\|(H-I)z\|^2 = n\lambda^2 z^\top (K+n\lambda I)^{-2} z \\ \text{variance}(K) &= \tfrac{1}{n}\operatorname{tr} CH^2 = \tfrac{1}{n}\operatorname{tr} CK^2(K+n\lambda I)^{-2}\end{aligned}$$

- When $C = \sigma^2 I$, $\text{variance}(K) = \frac{\sigma^2}{n}\operatorname{tr} H^2 = \frac{\sigma^2}{n}\operatorname{tr} K^2(K+n\lambda I)^{-2}$

- Degrees of freedom: $\operatorname{tr} K^2(K+n\lambda I)^{-2}$ or $\operatorname{tr} K(K+n\lambda I)^{-1}$
  - Implicit number of param. of smoothing mat. $H = K(K+n\lambda I)^{-1}$
  - Equal to p, if $\text{rank}(K) = p$ and $\lambda = 0$

- **Definition**: maximal marginal degrees of freedom

$$d = n\big\|\operatorname{diag}(H)\big\|_\infty = n\big\|\operatorname{diag}\big(K(K+n\lambda I)^{-1}\big)\big\|_\infty$$

Note: $\operatorname{tr} H^2 \leqslant \operatorname{tr} H = \big\|\operatorname{diag}(H))\big\|_1 \leqslant n\big\|\operatorname{diag}(H))\big\|_\infty = d$

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

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
- We "must" have $p \geqslant d$, if

(a) column sampling approximation obtained from held out data
(b) generalization error optimal

- **Does $p = O(d)$ suffice?**

Generalization performance of column sampling (Bach, 2012)

- **Assumptions**
  - $z \in \mathbb{R}^n$, $K \in \mathbb{R}^{n \times n}$ positive semidefinite, $\lambda > 0$,
  - $d = n \Big\| \operatorname{diag}\big(K(K + n\lambda I)^{-1}\big) \Big\|_\infty$ and $R^2 = \| \operatorname{diag}(K) \|_\infty$
  - $\varepsilon \in \mathbb{R}^n$ random vector with finite variance and zero mean
  - I uniform random subset of p indices in $\{1, \dots, n\}$
  - Column sampling approximation $L = K(V, I)K(I, I)^{-1}K(I, V)$
  - Estimate $\hat{z}_K = (K{+}n\lambda I)^{-1}K(z{+}\varepsilon)$ and $\hat{z}_L = (L{+}n\lambda I)^{-1}L(z{+}\varepsilon)$

Generalization performance of column sampling (Bach, 2012)

- **Assumptions**
  - $z \in \mathbb{R}^n$, $K \in \mathbb{R}^{n \times n}$ positive semidefinite, $\lambda > 0$,
  - $d = n \Big\| \operatorname{diag}\big(K(K + n\lambda I)^{-1}\big) \Big\|_\infty$ and $R^2 = \| \operatorname{diag}(K) \|_\infty$
  - $\varepsilon \in \mathbb{R}^n$ random vector with finite variance and zero mean
  - I uniform random subset of p indices in $\{1, \dots, n\}$
  - Column sampling approximation $L = K(V, I) K(I, I)^{-1} K(I, V)$
  - Estimate $\hat{z}_K = (K + n\lambda I)^{-1} K(z + \varepsilon)$ and $\hat{z}_L = (L + n\lambda I)^{-1} L(z + \varepsilon)$

- For any $\delta \in (0, 1)$, if $p \geqslant \big(\frac{32d}{\delta} + 2\big) \log \frac{nR^2}{\delta\lambda}$, then

$$\frac{1}{n} \mathbb{E}_I \mathbb{E}_\varepsilon \| \hat{z}_L - z \|^2 \leqslant \frac{1}{n} (1 + 4\delta) \mathbb{E}_\varepsilon \| \hat{z}_K - z \|^2.$$

Generalization performance of column sampling

- For any $\delta \in (0,1)$, if $p \geqslant \big(\frac{32d}{\delta} + 2\big) \log \frac{nR^2}{\delta\lambda}$, then

$$\frac{1}{n}\mathbb{E}_I\mathbb{E}_\varepsilon \|\hat{z}_L - z\|^2 \leqslant \frac{1}{n}(1+4\delta)\mathbb{E}_\varepsilon\|\hat{z}_K - z\|^2.$$

- **Discussion**
  - Proof technique: approximation of subsampled covariance matrices (Tropp, 2011; Gittens, 2011)
  - No assumptions on eigengap or on the noise
  - Relative approximation guarantee
  - Expectations, both with respect to the data (i.e., $\mathbb{E}_\varepsilon$) and the sampling of columns (i.e., $\mathbb{E}_I$)
  - Different from good approximation of K
  - Sufficient lower-bound for required rank p
  - Logarithmic term in λ

Beyond least-square regression
Self-concordant analysis of logistic regression

- Logistic loss $\ell(u) = \log(1 + e^{-u})$
  - No closed-form expressions

- **Self-concordance** (Nesterov and Nemirovski, 1994)
  - $g : \mathbb{R} \to \mathbb{R}$ is self-concordant iff $\forall u \in \mathbb{R}$, $|g'''(u)| \leqslant 2g''(u)^{3/2}$

- **Extension for logistic loss** (Bach, 2010): $\forall u \in \mathbb{R}$, $|g'''(u)| \leqslant g''(u)$

- Allows non-asymptotic analysis of logistic regression
  - With exact first-order term
  - Replace covariance by Fisher information matrix

Optimal choice of the regularization parameter λ

- **Eigenvalues of** $K = \Theta(n\mu_i)$, $i = 1, \dots, n$, 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	$n^{-1/2}$	$\log n/n$	$\log n$	$\kappa > 2\rho$
$e^{-\rho i}$	$e^{-\kappa i}$	$n^{-1}\log\frac{1}{\lambda}$	$\lambda^{\kappa/\rho}$	$n^{-\rho/\kappa}$	$\log n/n$	$\log n$	$\kappa < 2\rho$

- Always assume $\delta > 1/2$, $\beta > 1/2$, $\rho > 0$, $\kappa > 0$

Optimal choice of the regularization parameter λ

(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	$n^{-1/2}$	$\log n/n$	$\log n$	$\kappa > 2\rho$
$e^{-\rho i}$	$e^{-\kappa i}$	$n^{-1}\log\frac{1}{\lambda}$	$\lambda^{\kappa/\rho}$	$n^{-\rho/\kappa}$	$\log n/n$	$\log n$	$\kappa < 2\rho$

- Best possible performance (Johnstone, 1994; Steinwart et al., 2009)
  - if $\nu_i = O(i^{-2\delta})$: $O(n^{1/2\delta-1})$
  - if $\nu_i = O(e^{-\kappa i})$: $O(\log n/n)$
- Faster decay of components (ν_i) of $K \approx$ smoother functions
- Faster decay of eigenvalues (μ_i) of $K \approx$ smaller feature space
  - Overfitting if feature space too large
  - Numerical problems if feature space too small

Optimization algorithms with column sampling
Twice-differentiable losses

- Given rank p and regularization parameter λ

  1. Select at random p columns of K (without replacement)
  2. Compute $\Phi \in \mathbb{R}^{n \times p}$ such that $\Phi\Phi^\top = K(V,I)K(I,I)^{-1}K(I,V)$ using incomplete Cholesky decomposition
  3. Minimize $\min_{w \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, (\Phi w)_i) + \frac{\lambda}{2}\|w\|^2$ using Newton's method (i.e., a single linear system for the square loss).

- Complexity $O(p^2 n) \approx O(d^2 n)$

- Robustness to ill-conditioning and in particular to small values of λ

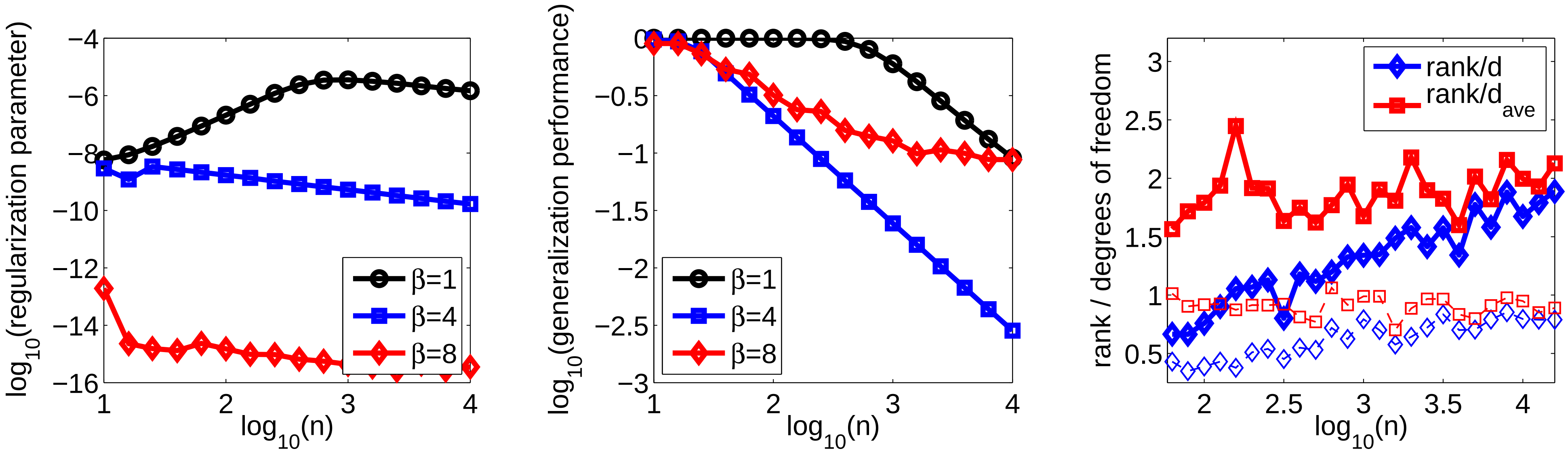
- Choice of p in practice?

Simulations on synthetic examples

- Periodic smoothing splines on $[0,1]$ and points $x_1,\dots,x_n$ uniformly spread over $[0,1]$

- $k(x,y)=\sum_{i=1}^{\infty} 2\mu_i \cos 2i\pi(x-y)$, and $f(x)=\sum_{i=1}^{\infty} 2\nu_i^{1/2} \cos 2i\pi x$

- $\nu_i = i^{-2\delta}$, $\mu_i = i^{-2\beta}$, $\delta=8$, $\beta=1,4,8$

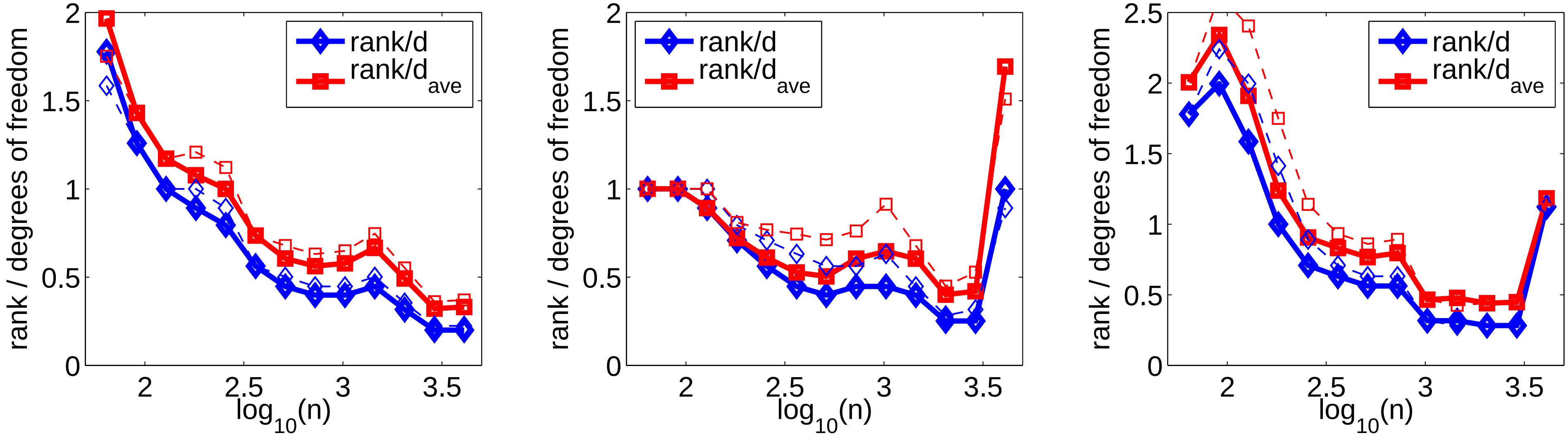


– *Left*: regularization parameter λ, *right*: predictive performance

– *Right*: sufficient rank to obtain 1% worse predictive performance

Simulations on *pumadyn* datasets

- Sufficient rank to obtain 1% worse predictive performance, over the degrees of freedom



- From left to right: *pumadyn* datasets *32fh*, *32nh*, *32nm*

Conclusions

- **Analysis of column sampling for kernel least-squares regression**
  - Degrees of freedom: both statistical and computational roles
- **Extensions**
  - Beyond uniform sampling (Boutsidis et al., 2009; S. Kumar, 2012)
  - Random design using results from Hsu et al. (2011)
  - Achieve $O(dn)$ running-time complexity
  - Beyond least-squares regression, e.g., logistic regression (Bach, 2010), SVM (Blanchard et al., 2008)
  - Online setting with properly decaying regularization parameter
  - Relationship with averaged stochastic gradient (Polyak and Juditsky, 1992)

References

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010. ISSN 1935-7524.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. *arXiv preprint arXiv:1208.2015*, 2012.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36(2):489–531, 2008.

A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *The Journal of Machine Learning Research*, 6:1579–1619, 2005.

S. Boucheron and P. Massart. A high-dimensional wilks phenomenon. *Probability theory and related fields*, 150(3-4):405–433, 2011.

C. Boutsidis, M. W. Mahoney, and P. Drineas. An improved approximation algorithm for the column subset selection problem. In *Proc. SODA*, 2009.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007. ISSN 1615-3375.

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

O. Dekel, S. Shalev-Shwartz, and Y. Singer. The Forgetron: A kernel-based perceptron on a fixed budget. In *Adv. NIPS*, 2005.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *J. Mac. Learn. Res.*, 2:243–264, 2001.

A. Gittens. The spectral norm error of the naive Nyström extension. *Arxiv preprint arXiv:1110.5305*, 2011.

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

D. Hsu, S. M. Kakade, and T. Zhang. An analysis of random design linear regression. *arXiv preprint arXiv:1106.2363*, 2011.

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.

I. M. Johnstone. Minimax Bayes, asymptotic minimax and sparse wavelet priors. *Statistical Decision Theory and Related Topics*, pages 303–326, 1994.

G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applicat.*, 33:82–95, 1971.

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.

Y. Nesterov and A. Nemirovski. *Interior-point polynomial algorithms in convex programming*. SIAM studies in Applied Mathematics, 1994.

F. Orabona, J. Keshet, and B. Caputo. The Projectron: a bounded kernel-based perceptron. In *Proc. ICML*, 2008.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20:1177–1184, 2007.

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

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

A. J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proc. ICML*, 2000.

K. Sridharan, N. Srebro, and S. Shalev-Shwartz. Fast rates for regularized objectives. *Advances in Neural Information Processing Systems*, 22, 2008.

I. Steinwart, D. Hush, C. Scovel, et al. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, pages 1–46, 2011.

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

Z. Wang, K. Crammer, and S. Vucetic. Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale svm training. *Journal of Machine Learning Research*, 13:3103–3131, 2012.

C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Adv. NIPS*, 2001.