Less is More: Computational Regularization by Subsampling

Lorenzo Rosasco
University of Genova - Istituto Italiano di Tecnologia
Massachusetts Institute of Technology
lcsl.mit.edu

joint work with Alessandro Rudi, Raffaello Camoriano

Paris

Laboratory for Computational and Statistical Learning

A Starting Point

Classically:
Statistics and optimization **distinct steps** in algorithm design

Empirical process theory + Optimization

A Starting Point

Classically:

Statistics and optimization **distinct steps** in algorithm design

Empirical process theory + Optimization

Large Scale:

Consider **interplay** between statistics and optimization!

(Bottou, Bousquet '08)

A Starting Point

Classically:
Statistics and optimization **distinct steps** in algorithm design

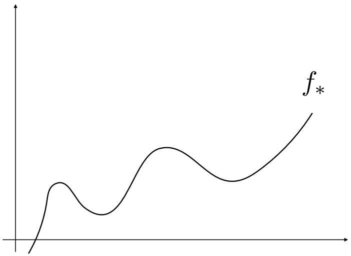
Empirical process theory + Optimization

Large Scale:
Consider **interplay** between statistics and optimization!
(Bottou, Bousquet '08)

Computational Regularization:
Computation "tricks"=regularization

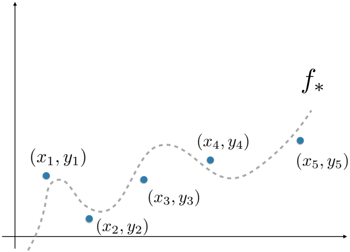
Supervised Learning

Problem: Estimate f^*



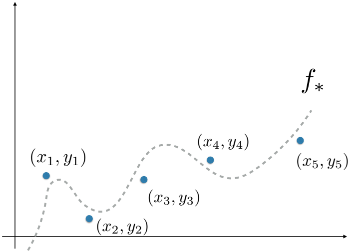
Supervised Learning

Problem: Estimate f^* given $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$



Supervised Learning

Problem: Estimate f^* given $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$



The Setting

$$y_i = f^*(x_i) + \varepsilon_i \qquad i \in \{1, \ldots, n\}$$

- $\varepsilon_i \in \mathbb{R}, x_i \in \mathbb{R}^d$ **random** (bounded but with unknown distribution)
- f^* **unknown**

Outline

Non-linear/non-parametric learning

$$\widehat{f}(x) \;=\; \sum_{i=1}^{M} c_i\, q(x, w_i)$$

Non-linear/non-parametric learning

$$\widehat{f}(x) \;=\; \sum_{i=1}^{M} c_i \, q(x, w_i)$$

- q non linear function

Non-linear/non-parametric learning

$$\widehat{f}(x) = \sum_{i=1}^{M} c_i\, q(x, w_i)$$

- q non linear function
- $w_i \in \mathbb{R}^d$ centers

Non-linear/non-parametric learning

$$\widehat{f}(x) = \sum_{i=1}^{M} c_i \, q(x, w_i)$$

- q non linear function
- $w_i \in \mathbb{R}^d$ centers
- $c_i \in \mathbb{R}$ coefficients

Non-linear/non-parametric learning

$$\widehat{f}(x) = \sum_{i=1}^{M} c_i \, q(x, w_i)$$

- q non linear function
- $w_i \in \mathbb{R}^d$ centers
- $c_i \in \mathbb{R}$ coefficients
- $M = M_n$ could/should *grow* with n

Non-linear/non-parametric learning

$$\widehat{f}(x) = \sum_{i=1}^{M} c_i \, q(x, w_i)$$

- q non linear function
- $w_i \in \mathbb{R}^d$ centers
- $c_i \in \mathbb{R}$ coefficients
- $M = M_n$ could/should *grow* with n

Question: How to choose w_i, c_i and M given S_n ?

Learning with Positive Definite Kernels

There is an *elegant* answer if:

- q is **symmetric**
- *all* the matrices $\widehat{Q}_{ij} = q(x_i, x_j)$ are **positive semi-definite**[1]

[1] They have non-negative eigenvalues

Learning with Positive Definite Kernels

There is an *elegant* answer if:

- q is **symmetric**
- *all* the matrices $\widehat{Q}_{ij} = q(x_i, x_j)$ are **positive semi-definite**[1]

Representer Theorem (Kimeldorf, Wahba '70; Schölkopf et al. '01)

- $M = n$,
- $w_i = x_i$,
- c_i by **convex** optimization!

[1] They have non-negative eigenvalues

Kernel Ridge Regression (KRR)

a.k.a. Tikhonov Regularization

$$\widehat{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|^2$$

where[2]

$$\mathcal{H} = \{ f \mid f(x) = \sum_{i=1}^{M} c_i q(x, w_i),\ c_i \in \mathbb{R}, \underbrace{w_i \in \mathbb{R}^d}_{\text{any center!}},\ \underbrace{\color{red}{M \in \mathbb{N}}}_{\text{any length!}} \}$$

[2] The norm is induced by the inner product $\langle f, f' \rangle = \sum_{i,j} c_i c_j' q(x_i, x_j)$

Kernel Ridge Regression (KRR)

a.k.a. Tikhonov Regularization

$$\widehat{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|^2$$

where[2]

$$\mathcal{H} = \{ f \mid f(x) = \sum_{i=1}^{M} c_i q(x, w_i),\ c_i \in \mathbb{R}, \underbrace{w_i \in \mathbb{R}^d}_{\text{any center!}},\ \underbrace{M \in \mathbb{N}}_{\text{any length!}} \}$$

Solution

$$\widehat{f}_\lambda = \sum_{i=1}^{n} c_i\, q(x, x_i) \quad \text{with} \quad c = (\widehat{Q} + \lambda n I)^{-1} \widehat{y}$$

[2] The norm is induced by the inner product $\langle f, f' \rangle = \sum_{i,j} c_i c'_j q(x_i, x_j)$

KRR: Statistics

KRR: Statistics

Well understood statistical properties:

Classical Theorem

If $f^* \in \mathcal{H}$, then

$$\lambda_* = \frac{1}{\sqrt{n}} \qquad \mathbb{E}\,(\widehat{f}_{\lambda_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

KRR: Statistics

Well understood statistical properties:

Classical Theorem

If $f^* \in \mathcal{H}$, then

$$\lambda_* = \frac{1}{\sqrt{n}} \qquad \mathbb{E}\,(\widehat{f}_{\lambda_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

Remarks

KRR: Statistics

Well understood statistical properties:

Classical Theorem

If $f^* \in \mathcal{H}$, then

$$\lambda_* = \frac{1}{\sqrt{n}} \qquad \mathbb{E}\,(\widehat{f}_{\lambda_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

Remarks

1. **Optimal nonparametric bound**

KRR: Statistics

Well understood statistical properties:

Classical Theorem

If $f^* \in \mathcal{H}$, then

$$\lambda_* = \frac{1}{\sqrt{n}} \qquad \mathbb{E}\,(\widehat{f}_{\lambda_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

Remarks

1. **Optimal nonparametric bound**
2. More refined results for smooth kernels

$$\lambda_* = n^{-\frac{1}{2s+1}}, \qquad \mathbb{E}\,(\widehat{f}_{\lambda_*}(x) - f^*(x))^2 \;\lesssim\; n^{-\frac{2s}{2s+1}}$$

KRR: Statistics

Well understood statistical properties:

Classical Theorem

If $f^* \in \mathcal{H}$, then

$$\lambda_* = \frac{1}{\sqrt{n}} \qquad \mathbb{E}\,(\widehat{f}_{\lambda_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

Remarks

1. **Optimal nonparametric bound**
2. More refined results for smooth kernels

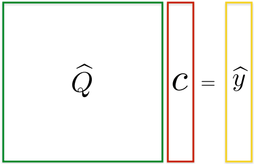
$$\lambda_* = n^{-\frac{1}{2s+1}}, \qquad \mathbb{E}\,(\widehat{f}_{\lambda_*}(x) - f^*(x))^2 \;\lesssim\; n^{-\frac{2s}{2s+1}}$$

3. **Adaptive tuning**, e.g. via cross validation
4. **Proofs**: inverse problems results + random matrices
   (Smale and Zhou + Caponnetto, De Vito, R.)

KRR: Optimization

$$\widehat{f}_\lambda = \sum_{i=1}^{n} c_i \, q(x, x_i) \quad \text{with} \quad c = (\widehat{Q} + \lambda n I)^{-1} \widehat{y}$$

Linear System



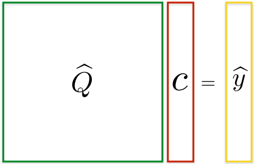
Complexity

- **Space** $O(n^2)$
- **Time** $O(n^3)$

KRR: Optimization

$$\widehat{f}_\lambda = \sum_{i=1}^{n} c_i \, q(x, x_i) \quad \text{with} \quad c = (\widehat{Q} + \lambda n I)^{-1} \widehat{y}$$

Linear System



Complexity

- **Space** $O(n^2)$
- **Time** $O(n^3)$

BIG DATA?

Running out of time and space ...

Can this be fixed?

Beyond Tikhonov: Spectral Filtering

$(\hat{Q} + \lambda I)^{-1}$ approximation of $\hat{Q}^{\dagger}$ controlled by λ

Beyond Tikhonov: Spectral Filtering

$(\hat{Q} + \lambda I)^{-1}$ approximation of $\hat{Q}^\dagger$ controlled by λ

Can we approximate $\hat{Q}^\dagger$ **by** saving computations?

Beyond Tikhonov: Spectral Filtering

$(\hat{Q} + \lambda I)^{-1}$ approximation of $\hat{Q}^\dagger$ controlled by λ

Can we approximate $\hat{Q}^\dagger$ **by** saving computations?

Yes!

Beyond Tikhonov: Spectral Filtering

$(\hat{Q} + \lambda I)^{-1}$ approximation of $\hat{Q}^\dagger$ controlled by λ

Can we approximate $\hat{Q}^\dagger$ **by** saving computations?

Yes!

Spectral filtering (Engl '96- inverse problems, Rosasco et al. 05- ML)

$$g_\lambda(\hat{Q}) \sim \hat{Q}^\dagger$$

The filter function g_λ defines the form of the approximation

Spectral filtering

Examples

- Tikhonov- ridge regression
- Truncated SVD– principal component regression
- Landweber iteration– GD/ L_2-boosting
- nu-method– accelerated GD/Chebyshev method
- ...

Spectral filtering

Examples

- Tikhonov- ridge regression
- Truncated SVD– principal component regression
- Landweber iteration– GD/ L_2-boosting
- nu-method– accelerated GD/Chebyshev method
- ...

Landweber iteration (truncated power series)...

$$c_t = g_t(\hat{Q}) = \gamma \sum_{r=0}^{t-1} (I - \gamma \hat{Q})^r \widehat{y}$$

Spectral filtering

Examples

- Tikhonov- ridge regression
- Truncated SVD– principal component regression
- Landweber iteration– GD/ L_2-boosting
- nu-method– accelerated GD/Chebyshev method
- ...

Landweber iteration (truncated power series)...

$$c_t = g_t(\hat{Q}) = \gamma \sum_{r=0}^{t-1} (I - \gamma \hat{Q})^r \widehat{y}$$

...it's GD for ERM!!

$$r = 1 \dots t \quad c_r = c_{r-1} - \gamma(\hat{Q} c_{r-1} - \hat{y}), \quad c_0 = 0$$

Statistics and computations with spectral filtering

The different filters achieve *essentially* **the same** optimal statistical error!

Statistics and computations with spectral filtering

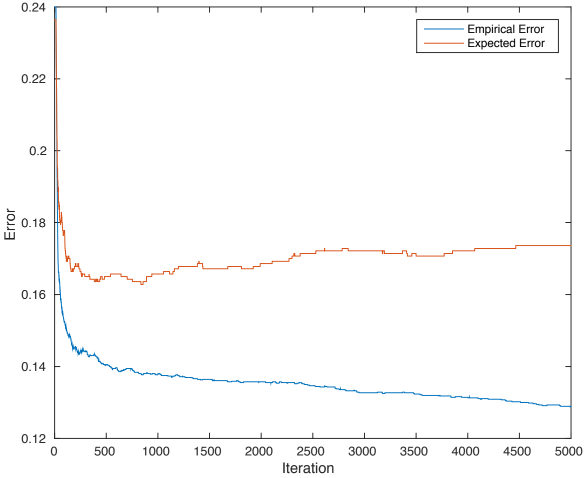
The different filters achieve *essentially* **the same** optimal statistical error!

Difference is in computations

Filter	*Time*	*Space*
Tikhonov	n^3	n^2
GD	$n^2\lambda_*^{-1}$	n^2
Accelerated GD	$n^2\lambda_*^{-1/2}$	n^2
Truncated SVD	$n^2\lambda_*^{-\gamma}$	n^2

Notet: $\lambda_*^{-1} = t$, for iterative methods

Semiconvergence



- Iterations control statistics **and** time complexity

Computational Regularization

Computational Regularization

BIG DATA?

Running out of ~~time and~~ space ...

Computational Regularization

BIG DATA?

Running out of ~~time and~~ space ...

Is there a principle to control statistics, time **and space** complexity?

Outline

Nonparametric Learning

Data Dependent Subsampling

Data Independent Subsampling

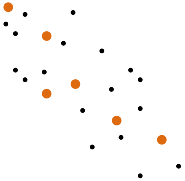
Subsampling

1. pick w_i at random...

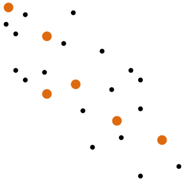
Subsampling

1. pick w_i at random... from training set
   (Smola, Schölkopf '00)

$$\tilde{w}_1, \ldots, \tilde{w}_M \subset x_1, \ldots x_n \qquad M \ll n$$



Subsampling



1. pick w_i at random... from training set
   (Smola, Scholköpf '00)

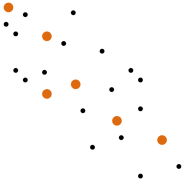
$$\tilde{w}_1, \ldots, \tilde{w}_M \subset x_1, \ldots x_n \qquad M \ll n$$

2. perform KRR on

$$\mathcal{H}_M = \{f \mid f(x) = \sum_{i=1}^{M} c_i q(x, \tilde{w}_i),\ c_i \in \mathbb{R}, \cancel{w_i \in \mathbb{R}^d}, \cancel{M \in \mathbb{N}}\}.$$

Subsampling

1. pick w_i at random... from training set
   (Smola, Schölkopf '00)

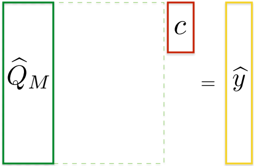


$$\tilde{w}_1, \ldots, \tilde{w}_M \subset x_1, \ldots x_n \qquad M \ll n$$

2. perform KRR on

$$\mathcal{H}_M = \{f \mid f(x) = \sum_{i=1}^{M} c_i q(x, \tilde{w}_i),\ c_i \in \mathbb{R}, \cancel{w_i \in \mathbb{R}^d}, \cancel{M \in \mathbb{N}}\}.$$

Linear System



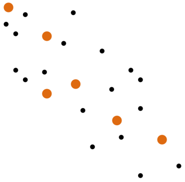
Complexity

- **Space** $\cancel{O(n^2)} \to O(nM)$
- **Time** $\cancel{O(n^3)} \to O(nM^2)$

Subsampling

1. pick w_i at random... from training set
   (Smola, Schölkopf '00)

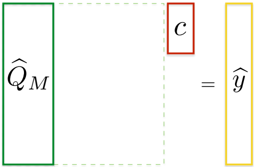
$$\tilde{w}_1, \dots, \tilde{w}_M \subset x_1, \dots x_n \qquad M \ll n$$



2. perform KRR on

$$\mathcal{H}_M = \{f \mid f(x) = \sum_{i=1}^{M} c_i q(x, \tilde{w}_i),\ c_i \in \mathbb{R}, \cancel{w_i \in \mathbb{R}^d}, \cancel{M \in \mathbb{N}}\}.$$

Linear System



Complexity

- **Space** $\cancel{O(n^2)} \to O(nM)$
- **Time** $\cancel{O(n^3)} \to O(nM^2)$

What about statistics? What's the price for efficient computations?

Putting our Result in Context

- *Many* **different subsampling** schemes
  (Smola, Scholkopf '00; Williams, Seeger '01; ... 20+)

Putting our Result in Context

- *Many* **different subsampling** schemes
  (Smola, Scholkopf '00; Williams, Seeger '01; ... 20+)

- **Theoretical guarantees** mainly on **matrix approximation**
  (Mahoney and Drineas '09; Cortes et al '10, Kumar et al.'12 ... 10+)

$$\|\widehat{Q} - \widetilde{Q}_M\| \lesssim \frac{1}{\sqrt{M}}$$

Putting our Result in Context

- *Many* **different subsampling** schemes
  (Smola, Scholkopf '00; Williams, Seeger '01; . . . 20+)

- **Theoretical guarantees** mainly on **matrix approximation**
  (Mahoney and Drineas '09; Cortes et al '10, Kumar et al.'12 . . . 10+)

$$\|\widehat{Q} - \widetilde{Q}_M\| \lesssim \frac{1}{\sqrt{M}}$$

- Statistical guarantees **suboptimal** or in **restricted setting**
  (Cortes et al. '10; Jin et al. '11, Bach '13, Alaoui, Mahoney '14)

Main Result

(Rudi, Camoriano, Rosasco, '15)

Theorem

If $f^* \in \mathcal{H}$, then

$$\lambda_* = \frac{1}{\sqrt{n}} \quad , M_* = \frac{1}{\lambda_*}, \quad \mathbb{E}\,(\widehat{f}_{\lambda_*, M_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

Main Result

(Rudi, Camoriano, Rosasco, '15)

Theorem

If $f^* \in \mathcal{H}$, then

$$\lambda_* = \frac{1}{\sqrt{n}} \quad , M_* = \frac{1}{\lambda_*}, \quad \mathbb{E}\,(\widehat{f}_{\lambda_*, M_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

Remarks

Main Result

(Rudi, Camoriano, Rosasco, '15)

Theorem

If $f^* \in \mathcal{H}$*, then*

$$\lambda_* = \frac{1}{\sqrt{n}} \quad , M_* = \frac{1}{\lambda_*}, \quad \mathbb{E}\,(\widehat{f}_{\lambda_*, M_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

Remarks

1. Subsampling achives **optimal** bound...

Main Result

(Rudi, Camoriano, Rosasco, '15)

Theorem

If $f^* \in \mathcal{H}$, then

$$\lambda_* = \frac{1}{\sqrt{n}} \quad , M_* = \frac{1}{\lambda_*}, \quad \mathbb{E}\,(\widehat{f}_{\lambda_*, M_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

Remarks

1. Subsampling achives **optimal** bound...
2. ...with $M_* \sim \sqrt{n}$!!

Main Result

(Rudi, Camoriano, Rosasco, '15)

Theorem

If $f^* \in \mathcal{H}$, then

$$\lambda_* = \frac{1}{\sqrt{n}} \quad , M_* = \frac{1}{\lambda_*}, \quad \mathbb{E}\,(\widehat{f}_{\lambda_*, M_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

Remarks

1. Subsampling achives **optimal** bound...
2. ...with $M_* \sim \sqrt{n}$!!
3. **More generally**,

$$\lambda_* = n^{-\frac{1}{2s+1}}, \quad M_* = \frac{1}{\lambda_*}, \quad \mathbb{E}_x\,(\widehat{f}_{\lambda_*, M_*}(x) - f^*(x))^2 \;\lesssim\; n^{-\frac{2s}{2s+1}}$$

Main Result

(Rudi, Camoriano, Rosasco, '15)

Theorem

If $f^* \in \mathcal{H}$, then

$$\lambda_* = \frac{1}{\sqrt{n}} \quad , M_* = \frac{1}{\lambda_*}, \quad \mathbb{E}\,(\widehat{f}_{\lambda_*, M_*}(x) - f^*(x))^2 \;\lesssim\; \frac{1}{\sqrt{n}}$$

Remarks

1. Subsampling achives **optimal** bound...
2. ...with $M_* \sim \sqrt{n}$!!
3. **More generally**,

$$\lambda_* = n^{-\frac{1}{2s+1}}, \quad M_* = \frac{1}{\lambda_*}, \quad \mathbb{E}_x\,(\widehat{f}_{\lambda_*, M_*}(x) - f^*(x))^2 \;\lesssim\; n^{-\frac{2s}{2s+1}}$$

Note: An interesting insight is obtained rewriting the result...

Computational Regularization by Subsampling

(Rudi, Camoriano, Rosasco, '15)

A simple idea: "*swap*" the role of λ and M...

Computational Regularization by Subsampling

(Rudi, Camoriano, Rosasco, '15)

A simple idea: *"swap"* the role of λ and M...

Theorem

If $f^* \in \mathcal{H}$ with a smooth kernel, then

$$M_* = n^{\frac{1}{2s+1}}, \quad \lambda_* = \frac{1}{M_*}, \quad \mathbb{E}_x\,(\widehat{f}_{\lambda_*,M_*}(x) - f^*(x))^2 \;\lesssim\; n^{-\frac{2s}{2s+1}}$$

Computational Regularization by Subsampling

(Rudi, Camoriano, Rosasco, '15)

A simple idea: *"swap"* the role of λ and M...

Theorem

If $f^* \in \mathcal{H}$ with a smooth kernel, then

$$M_* = n^{\frac{1}{2s+1}}, \quad \lambda_* = \frac{1}{M_*}, \quad \mathbb{E}_x \left(\widehat{f}_{\lambda_*, M_*}(x) - f^*(x)\right)^2 \lesssim n^{-\frac{2s}{2s+1}}$$

- λ and M play the same role...
  ...new interpretation: **subsampling regularizes**!

Computational Regularization by Subsampling

(Rudi, Camoriano, Rosasco, '15)

A simple idea: *"swap"* the role of λ and M...

Theorem

If $f^* \in \mathcal{H}$ with a smooth kernel, then

$$M_* = n^{\frac{1}{2s+1}}, \quad \lambda_* = \frac{1}{M_*}, \quad \mathbb{E}_x\,(\widehat{f}_{\lambda_*,M_*}(x) - f^*(x))^2 \;\lesssim\; n^{-\frac{2s}{2s+1}}$$

- λ and M play the same role...
  ...new interpretation: **subsampling regularizes**!
- New natural **incremental** algorithm...

Algorithm

Computational Regularization by Subsampling

(Rudi, Camoriano, Rosasco, '15)

A simple idea: *"swap"* the role of λ and M...

Theorem

If $f^* \in \mathcal{H}$ with a smooth kernel, then

$$M_* = n^{\frac{1}{2s+1}}, \quad \lambda_* = \frac{1}{M_*}, \quad \mathbb{E}_x\,(\widehat{f}_{\lambda_*,M_*}(x) - f^*(x))^2 \lesssim n^{-\frac{2s}{2s+1}}$$

- λ and M play the same role...
  ...new interpretation: **subsampling regularizes**!
- New natural **incremental** algorithm...

Algorithm

1. *Pick a center + compute solution*

Computational Regularization by Subsampling

(Rudi, Camoriano, Rosasco, '15)

A simple idea: *"swap"* the role of λ and M...

Theorem

If $f^* \in \mathcal{H}$ with a smooth kernel, then

$$M_* = n^{\frac{1}{2s+1}}, \quad \lambda_* = \frac{1}{M_*}, \quad \mathbb{E}_x\,(\widehat{f}_{\lambda_*,M_*}(x) - f^*(x))^2 \lesssim n^{-\frac{2s}{2s+1}}$$

- λ and M play the same role...
  ...new interpretation: **subsampling regularizes**!
- New natural **incremental** algorithm...

Algorithm

1. *Pick a center + compute solution*
2. *Pick another center +* **rank one update**

Computational Regularization by Subsampling

(Rudi, Camoriano, Rosasco, '15)

A simple idea: *"swap"* the role of λ and M...

Theorem

If $f^* \in \mathcal{H}$ with a smooth kernel, then

$$M_* = n^{\frac{1}{2s+1}}, \quad \lambda_* = \frac{1}{M_*}, \quad \mathbb{E}_x\,(\widehat{f}_{\lambda_*,M_*}(x) - f^*(x))^2 \lesssim n^{-\frac{2s}{2s+1}}$$

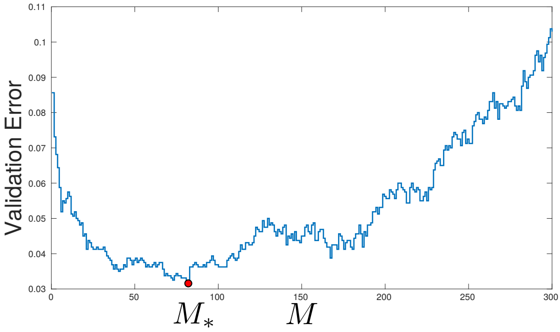
- λ and M play the same role...
  ...new interpretation: **subsampling regularizes**!
- New natural **incremental** algorithm...

Algorithm

1. *Pick a center + compute solution*
2. *Pick another center +* **rank one update**
3. *Pick another center ...*

Nÿstrom CoRe Illustrated

n, λ are fixed



Computation controls stability!

Time/space requirement tailored to **generalization**

Experiments

comparable/better w.r.t. the state of the art

Dataset	n_{tr}	d	*Incremental CoRe*	*Standard KRLS*	*Standard Nyström*	*Random Features*	*Fastfood RF*
Ins. Co.	5822	85	$0.23180 \pm 4 \times 10^{-5}$	**0.231**	0.232	0.266	0.264
CPU	6554	21	$\mathbf{2.8466 \pm 0.0497}$	7.271	6.758	7.103	7.366
CT slices	42800	384	$\mathbf{7.1106 \pm 0.0772}$	NA	60.683	49.491	43.858
Year Pred.	463715	90	$\mathbf{0.10470 \pm 5 \times 10^{-5}}$	NA	0.113	0.123	0.115
Forest	522910	54	0.9638 ± 0.0186	NA	**0.837**	0.840	0.840

- Random Features (Rahimi, Recht '07)
- Fastfood (Le et al. '13)

Summary so far

- **Optimal** learning with data dependent subsampling
- Computational regularization: subsampling **regularizes**!

Summary so far

- **Optimal** learning with data dependent subsampling
- Computational regularization: subsampling **regularizes**!

Few more questions:

- Can one do better than **uniform** sampling?

Summary so far

- **Optimal** learning with data dependent subsampling
- Computational regularization: subsampling **regularizes**!

Few more questions:

- Can one do better than **uniform** sampling?
  Yes: leverage score sampling...
- What about **data independent** sampling?

Outline

Nonparametric Learning

Data Dependent Subsampling

Data Independent Subsampling

Random Features

$$\widehat{f}(x) \;=\; \sum_{i=1}^{M} c_i \, q(x, w_i)$$

Random Features

$$\widehat{f}(x) = \sum_{i=1}^{M} c_i \, q(x, w_i)$$

- q general non linear function

Random Features

$$\widehat{f}(x) = \sum_{i=1}^{M} c_i \, q(x, w_i)$$

- q general non linear function
- pick $\tilde{w}_i$ at random **according to a distribution** μ

$$\tilde{w}_1, \ldots, \tilde{w}_M \sim \mu$$

Random Features

$$\widehat{f}(x) \;=\; \sum_{i=1}^{M} c_i\, q(x, w_i)$$

- q general non linear function
- pick $\tilde{w}_i$ at random **according to a distribution** μ

$$\tilde{w}_1, \dots, \tilde{w}_M \sim \mu$$

- perform KRR on

$$\mathcal{H}_M = \{f \mid f(x) = \sum_{i=1}^{M} c_i q(x, \tilde{w}_i),\; c_i \in \mathbb{R}\}.$$

Random Fourier Features

(Rahimi, Recht '07)

Consider

$$q(x, w) = e^{iw^T x},$$

Random Fourier Features

(Rahimi, Recht '07)

Consider

$$q(x, w) = e^{iw^T x}, \qquad w \sim \mu(w) = \mathcal{N}(0, I)$$

Random Fourier Features

(Rahimi, Recht '07)

Consider

$$q(x, w) = e^{iw^T x}, \qquad w \sim \mu(w) = \mathcal{N}(0, I)$$

Then

$$\mathbb{E}_w \left[q(x, w) \overline{q(x', w)} \right] = e^{-\|x - x'\|^2 \gamma} = K(x, x')$$

Random Fourier Features

(Rahimi, Recht '07)

Consider

$$q(x, w) = e^{iw^T x}, \qquad\qquad w \sim \mu(w) = \mathcal{N}(0, I)$$

Then

$$\mathbb{E}_w \left[q(x, w) \overline{q(x', w)} \right] = e^{-\|x - x'\|^2 \gamma} = K(x, x')$$

By sampling $\tilde{w}_1, \dots, \tilde{w}_M$ we are considering the **approximating kernel**

$$\frac{1}{M} \sum_{i=1}^{M} \left[q(x, \tilde{w}_i) \overline{q(x', \tilde{w}_i)} \right] = \widetilde{K}_M(x, x')$$

More Random Features

- **translation invariant** kernels $K(x, x') = H(x - x')$,

$$q(x, w) = e^{iw^T x}, \qquad w \sim \mu = \mathcal{F}(H)$$

- infinite **neural nets** kernels

$$q(x, w) = |w^T x + b|_+, \qquad (w, b) \sim \mu = U[\mathbb{S}^d]$$

- infinite **dot product** kernels
- homogeneous **additive** kernels
- **group invariant** kernels
- ...

Note: Connections with **hashing** and **sketching** techniques.

Properties of Random Features

Properties of Random Features

Optimization

- Time: $\cancel{O(n^3)}$ $O(nM^2)$
- Space: $\cancel{O(n^2)}$ $O(nM)$

Properties of Random Features

Optimization

- Time: ~~$O(n^3)$~~ $O(nM^2)$
- Space: ~~$O(n^2)$~~ $O(nM)$

Statistics

As before: **do we pay a price for efficient computations?**

Previous works

Previous works

- ***Many*** different random features for different kernels
  (Rahimi, Recht '07, Vedaldi, Zisserman, . . . 10+)

Previous works

- ***Many*** different random features for different kernels
  (Rahimi, Recht '07, Vedaldi, Zisserman, . . . 10+)

- Theoretical guarantees: mainly **kernel approximation**
  (Rahimi, Recht '07, . . . , Sriperumbudur and Szabo '15)

$$|K(x, x') - \widetilde{K}_M(x, x')| \lesssim \frac{1}{\sqrt{M}}$$

Previous works

- ***Many*** different random features for different kernels
  (Rahimi, Recht '07, Vedaldi, Zisserman, ... 10+)

- Theoretical guarantees: mainly **kernel approximation**
  (Rahimi, Recht '07, ..., Sriperumbudur and Szabo '15)

$$|K(x, x') - \widetilde{K}_M(x, x')| \lesssim \frac{1}{\sqrt{M}}$$

- Statistical guarantees **suboptimal or in restricted setting**
  (Rahimi, Recht '09, Yang et al. '13 ...,Bach '15)

Main Result

Let

$$q(x, w) = e^{iw^T x},$$

Main Result

Let

$$q(x, w) = e^{iw^T x}, \qquad w \sim \mu(w) = c_d \left(\frac{1}{1 + \|w\|^2} \right)^{\frac{d+1}{2}}$$

Main Result

Let

$$q(x,w) = e^{iw^T x}, \qquad w \sim \mu(w) = c_d \left(\frac{1}{1+\|w\|^2}\right)^{\frac{d+1}{2}}$$

Theorem

If $f_* \in \mathcal{H}_s$ Sobolev space, then

$$\lambda_* = n^{-\frac{1}{2s+1}}, \quad M_* = \frac{1}{\lambda_*^{2s}}, \quad \mathbb{E}\,(\widehat{f}_{\lambda_*,M_*}(x) - f^*(x))^2 \;\lesssim\; n^{-\frac{2s}{2s+1}}$$

Main Result

Let

$$q(x,w) = e^{iw^T x}, \qquad w \sim \mu(w) = c_d \left(\frac{1}{1+\|w\|^2} \right)^{\frac{d+1}{2}}$$

Theorem

If $f_* \in \mathcal{H}_s$ Sobolev space, then

$$\lambda_* = n^{-\frac{1}{2s+1}}, \quad M_* = \frac{1}{\lambda_*^{2s}}, \quad \mathbb{E}\left(\widehat{f}_{\lambda_*,M_*}(x) - f^*(x)\right)^2 \lesssim n^{-\frac{2s}{2s+1}}$$

- Random features achieve **optimal** bound!

Main Result

Let

$$q(x,w) = e^{iw^T x}, \qquad w \sim \mu(w) = c_d \left(\frac{1}{1+\|w\|^2}\right)^{\frac{d+1}{2}}$$

Theorem

If $f_* \in \mathcal{H}_s$ Sobolev space, then

$$\lambda_* = n^{-\frac{1}{2s+1}}, \quad M_* = \frac{1}{\lambda_*^{2s}}, \quad \mathbb{E}\left(\widehat{f}_{\lambda_*,M_*}(x) - f^*(x)\right)^2 \lesssim n^{-\frac{2s}{2s+1}}$$

- Random features achieve **optimal** bound!
- Efficient worst case subsampling $M_* \sim \sqrt{n}$– but cannot exploit smoothness.

Remarks & Extensions

Nÿstrom vs Random features

- Both achieve optimal rates
- Nÿstrom seems to need fewer samples (random centers)

Remarks & Extensions

Nÿstrom vs Random features

- Both achieve optimal rates
- Nÿstrom seems to need fewer samples (random centers)

How tight are the results?

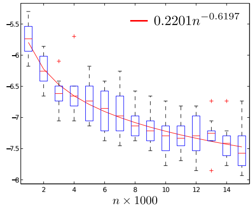
Remarks & Extensions

Nÿstrom vs Random features

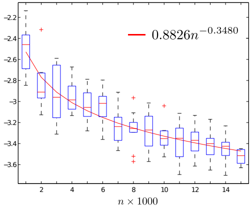
- Both achieve optimal rates
- Nÿstrom seems to need fewer samples (random centers)

How tight are the results?

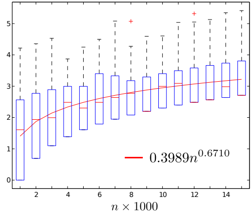
$\log \lambda$



Test Error



$\log M$



Contributions

- **Optimal bounds** for data dependent/independent subsampling
- Subsampling: Nÿstrom vs Random features
- Beyond ridge regression: **early stopping** and multiple passes SGD (see arxiv)

Contributions

- **Optimal bounds** for data dependent/independent subsampling
- Subsampling: Nÿstrom vs Random features
- Beyond ridge regression: **early stopping** and multiple passes SGD (see arxiv)

Some questions:

- Quest for the **best** sampling
- **Regularization by projection**: inverse problems and preconditioning
- Beyond randomization: **non convex neural nets optimization**?

Contributions

- **Optimal bounds** for data dependent/independent subsampling
- Subsampling: Nÿstrom vs Random features
- Beyond ridge regression: **early stopping** and multiple passes SGD (see arxiv)

Some questions:

- Quest for the **best** sampling
- **Regularization by projection**: inverse problems and preconditioning
- Beyond randomization: **non convex neural nets optimization**?

Some perspectives:

- **Computational regularization**: subsampling regularizes
- **Algorithm design**: control **stability** for good statistics/computations