

Stochastic Gradient Methods for non-parametric estimation

Noah Simon

University of Washington
Department of Biostatistics

March 20, 2019

Problem Setup

Suppose we have n observations ($i = 1, \dots, n$)

On each of which we measure features... $\mathbf{x}_i$

and a continuous outcome y_i .

Further suppose

$$y_i = f^*(\mathbf{x}_i) + \epsilon_i$$

for some iid, mean 0 errors ϵ_i .

Often of interest to estimate f^*

Informally The Question of Interest

Suppose I don't want to make strong assumptions on f^* ...

perhaps I just want to assume it has bounded derivatives

How do I estimate f^* in a statistically and computationally efficient way?

Informally The Question of Interest

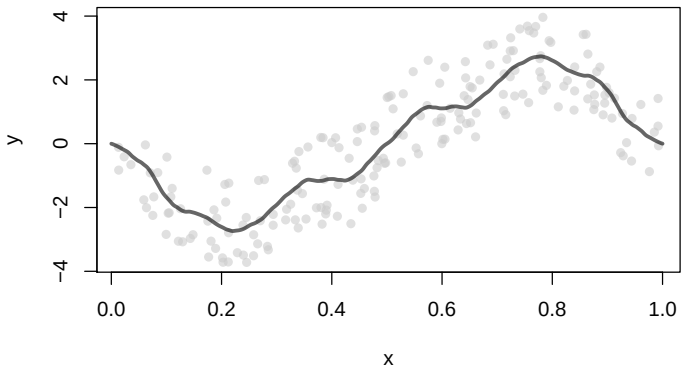
Let's make things a bit more concrete...

Suppose I **know** that

$$f^* \in \mathcal{F} \equiv \left\{ f \in \mathcal{L}_2 \mid \int_X [f''(x)]^2 dx < \infty \right\}$$

How do I estimate f^* from data?

Example



Outline of this talk

First cover two “classical” (offline) methods
and characterize statistical and computational performance

Then, we think about moving those ideas online
and talk about challenges to clear.

Failure of empirical risk minimization

The simplest approach would be to solve

$$\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

However, this fails!

Because we will interpolate the data.

Statistically “efficient” approaches

Two primary approaches are

- Projection Estimators
- Penalized Estimators

Projection Estimators

Here, we “diagonalize” $\mathcal{F}$:

Find functions $\psi_1, \dots, \psi_k, \dots \in \mathcal{F} \dots$

and increasing $\lambda_1, \dots, \lambda_k, \dots \in \mathbb{R}^+$

such that $\mathcal{F}$ can be written as

$$\mathcal{F} = \left\{ f = \sum_{j=1}^{\infty} \beta_j \psi_j \left| \sum_{j=1}^{\infty} \lambda_j \beta_j^2 < \infty \right. \right\}$$

Projection Estimators

Considering our example problem

$$\begin{aligned}\mathcal{F} &\equiv \left\{ f \in \mathcal{L}_2[0, 1] \left| \int_X [f''(x)]^2 dx < \infty \right. \right\} \\ &= \left\{ f = \sum_{j=1}^{\infty} \beta_j \psi_j \left| \sum_{j=1}^{\infty} \lambda_j \beta_j^2 < \infty \right. \right\}\end{aligned}$$

We have that

$$\psi_{2j}(x) = \sqrt{2} \cos(2\pi jx) \qquad \psi_{2j+1}(x) = \sqrt{2} \sin(2\pi jx)$$

with $\lambda_j = j^4$ (ish)

Projection Estimators

We now solve

$$\hat{\beta} \leftarrow \operatorname{argmin}_{\underline{\beta} \in \mathbb{R}^J} \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^J \beta_j \psi_j(x_i) \right)^2$$

with $\hat{f} \leftarrow \sum_{j=1}^J \hat{\beta}_j \psi_j$

where we choose $J \equiv J(n)$ strategically!

Projection Estimators

We now solve

$$\hat{\beta} \leftarrow \operatorname{argmin}_{\underline{\beta} \in \mathbb{R}^J} \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^J \beta_j \psi_j(x_i) \right)^2$$

with $\hat{f} \leftarrow \sum_{j=1}^J \hat{\beta}_j \psi_j$

By selecting $J < \infty$ we induce bias...

bias is on order of $\sum_{j=J+1}^{\infty} \lambda_j^2$

However, we reduce variance!

variance is on order of J

Projection Estimators

We now solve

$$\hat{\beta}_n \leftarrow \operatorname{argmin}_{\underline{\beta} \in \mathbb{R}^J} \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^J \beta_j \psi_j(x_i) \right)^2$$

Optimal $J(n)$ is $\sim n^{\frac{1}{2\alpha+1}}$ where $\alpha \sim \#$ of bdd derivatives of f^*

Resulting in statistically minimax rate-optimal estimator

$$\left\| \hat{f} - f^* \right\|^2 \sim n^{\frac{-2\alpha}{2\alpha+1}}$$

Projection Estimators

This is just a least-squares problem with $J(n)$ features

For optimal $J(n) \sim n^{\frac{1}{2\alpha+1}}$ ($\alpha = 2$ in example)

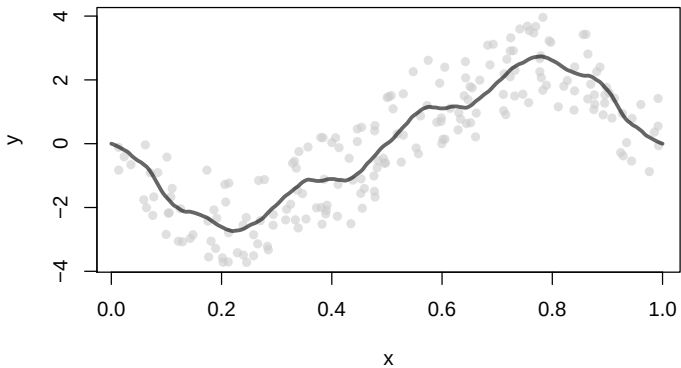
Computation time via IP

$$\sim n * \left(n^{\frac{1}{2\alpha+1}} \right)^2 = n^{\frac{2\alpha+3}{2\alpha+1}}$$

A single grad step takes

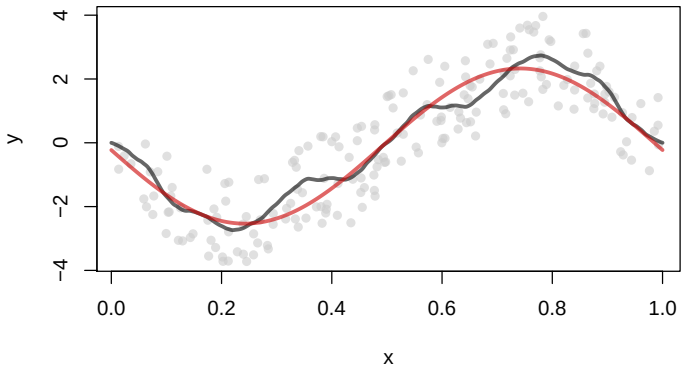
$$\sim n * \left(n^{\frac{1}{2\alpha+1}} \right) = n^{\frac{2\alpha+2}{2\alpha+1}}$$

Example



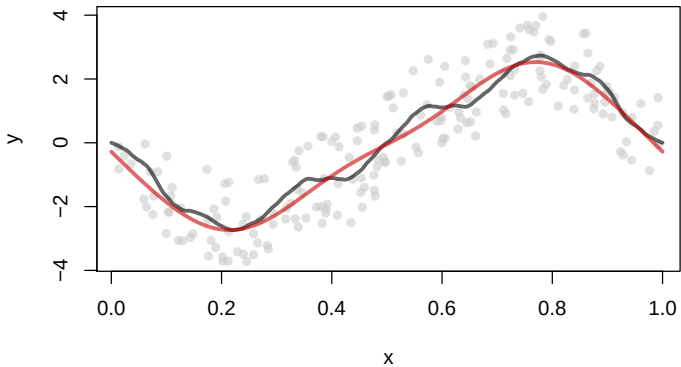
Example

Fit with $J = 3$



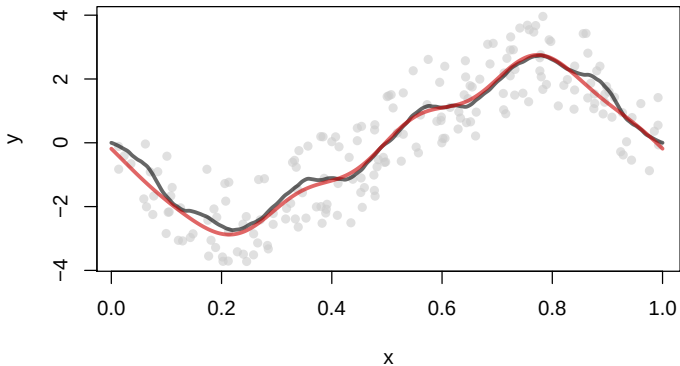
Example

Fit with $J = 7$



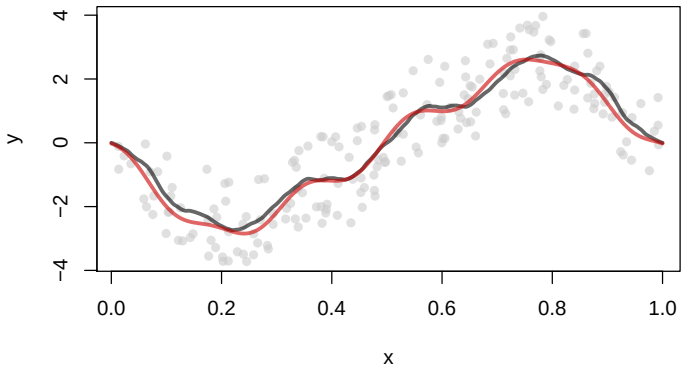
Example

Fit with $J = 11$



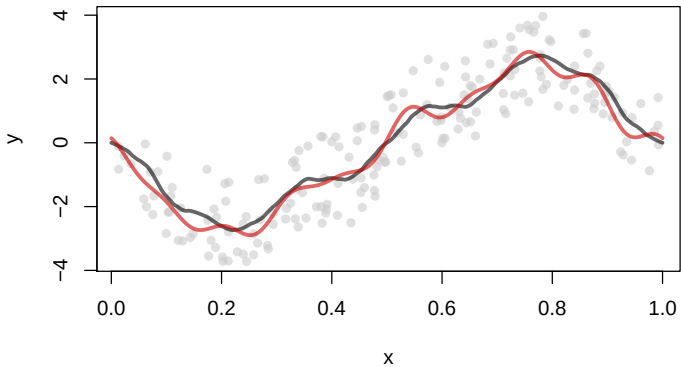
Example

Fit with $J = 15$



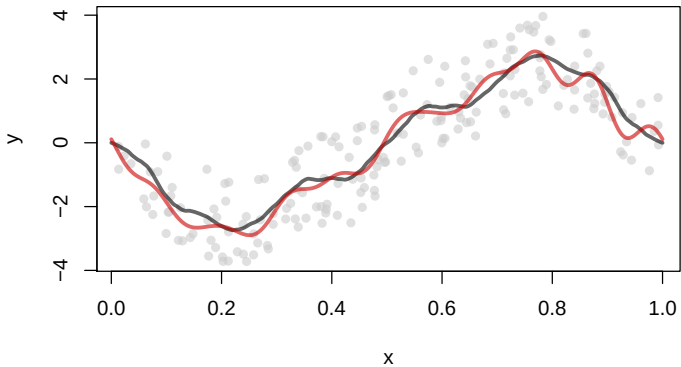
Example

Fit with $J = 19$



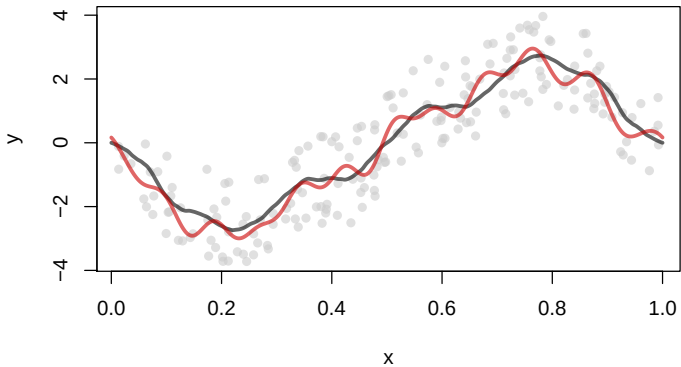
Example

Fit with $J = 23$



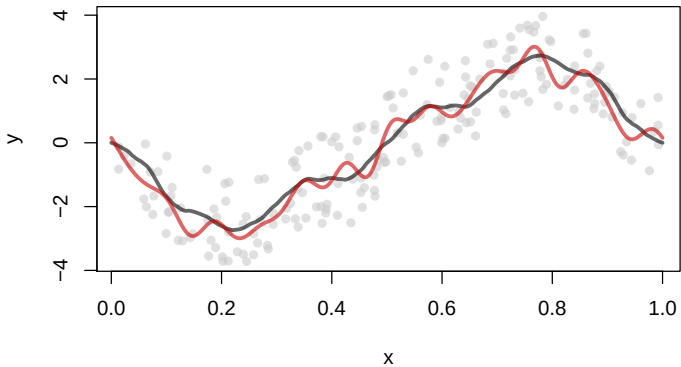
Example

Fit with $J = 27$



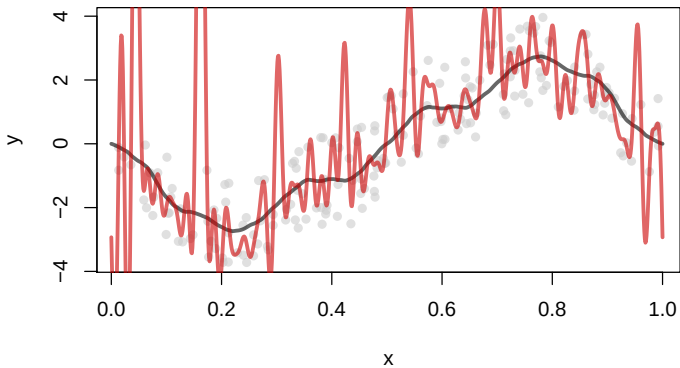
Example

Fit with $J = 31$



Example

Fit with $J = 101$



Statistically “efficient” approaches

Two primary approaches are

- Projection Estimators
- Penalized Estimators

Penalized Estimation

For $\mathcal{F}$ defined by

$$\mathcal{F}_P = \{f \in \mathcal{L}_2 \mid P(f) < \infty\}$$

Penalized estimator solves

$$\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \rho P(f)$$

Where $\rho \equiv \rho(n)$ strategically chosen.

Penalized Estimation

Penalized estimator solves

$$\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \rho P(f)$$

For $\rho \equiv \rho(n)$ strategically chosen.

This also results in a statistically rate-optimal estimator.

However, estimation here looks potentially hard?

Penalized Estimation

$$\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \rho P(f)$$

For some P there are (near) Linear-time, custom solvers:

$$(\text{eg. } P(f) = \int |f'(x)| dx, \int |f^{(k)}(x)|^2 dx)$$

For many P there are not even clear finite dimensional reductions.

Penalized Estimation

$$\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \rho P(f)$$

RKHS penalties form an **[un]**happy medium:

- There is an exact finite dimensional reduction...
- But it can be painful to solve!

Penalized Estimation

$$\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \rho P(f)$$

Let's return to

$$P(f) = \sum_{j=1}^{\infty} \lambda_j |\langle f, \psi_j \rangle|^2$$

Penalized Estimation

Let's return to

$$P(f) = \sum_{j=1}^{\infty} \lambda_j |\langle f, \psi_j \rangle|^2$$

To work here we define an inner-product:

$$\langle f, g \rangle_P = \sum_{j=1}^{\infty} \lambda_j \langle f, \psi_j \rangle \langle g, \psi_j \rangle$$

and a special function:

$$K(x, z) = \sum_{j=1}^{\infty} \lambda_j^{-1} \psi_j(x) \psi_j(z)$$

Exact Solutions in RKHS

To solve

$$\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \rho \langle f, f \rangle_P$$

Instead consider modified problem:

$$\hat{\beta} \leftarrow \operatorname{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^n \beta_j K(x_i, x_j) \right)^2 + \rho \sum_{i,j \leq n} \beta_i \beta_j K(x_i, x_j)$$

And have **exactly** $\hat{f} = \sum_{j=1}^n \hat{\beta}_j K(\cdot, x_j)$

Computational/Statistical Difficulty in RKHS

$$\hat{\beta} \leftarrow \operatorname{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{n} \|y - \mathbb{K}\beta\|_2^2 + \rho \|\mathbb{K}\beta\|_2^2$$

As noted these estimators will be **rate optimal**

Computational/Statistical Difficulty in RKHS

$$\hat{\beta} \leftarrow \operatorname{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{n} \|y - \mathbb{K}\beta\|_2^2 + \rho \|\mathbb{K}\beta\|_2^2$$

Computation time via (generic) IP

$$\sim n^3$$

A single grad step takes

$$\sim n^2$$

Computational/Statistical Difficulty in RKHS

$$\hat{\beta} \leftarrow \operatorname{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{n} \|y - \mathbb{K}\beta\|_2^2 + \rho \|\mathbb{K}\beta\|_2^2$$

Computation time via (generic) IP

$$\sim n^3$$

A single grad step takes

$$\sim n^2$$

ew..

Projection vs Penalized Problems

Our original example, $\int_X [f''(x)]^2$, lives in RKHS framework.

Both Penalized and Proj estimators are rate-optimal



Proj Estimator Computation:

- $\sim n^{7/5}$ for IP
- $\sim n^{6/5}$ for single grad



Penalized Estimator Computation:

- $\sim n^3$ for IP
- $\sim n^2$ for single grad



Projection vs Penalized Problems

Why such a big difference??

Statistical Optimality doesn't need **exact** penalized solution...

We are using too many basis elements (for computation)!

Projection vs Penalized Problems

Why such a big difference??

Statistical Optimality doesn't need **exact** penalized solution...

We are using too many basis elements (for computation)!

Projection Estimator considers J basis fcns

(diagonalization-based functions)

Kernel-based Estimator considers n basis fcns

(Kernel-based functions, $K(\cdot, x_i)$)

Online for better efficiency?

In RKHS spaces, if λ_j grow like $j^{2\alpha} \dots$

hoping we can get rate-optimal est in $n^{\frac{2\alpha+2}{2\alpha+1}}$ computation.

Ideally want this for every n sufficiently large.

(Don't want to prespecify a single $J(n)$)

Different views of stochastic optimization

May think of Stochastic Opt as method for solving an empirical minimization by sampling obs:

eg.,

$$\hat{\beta}_n \leftarrow \operatorname{argmin}_{\underline{\beta} \in \mathbb{R}^K} \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^K \beta_j \psi_j(x_i) \right)^2$$

In general, this is not how statisticians analyze it**.

Different views of stochastic optimization

Instead, think about solving

$$\begin{aligned} f^* &= \operatorname{argmin}_{f \in \mathcal{F}} E[(y - f(x))^2] \\ &= \operatorname{argmin}_{f \in \mathcal{F}} \int (y - f(x))^2 d\mathbb{P}(y, x) \end{aligned}$$

By sampling (x_i, y_i) pairs $(i = 1, \dots, p)$, iid from $\mathbb{P}$

and directly calculating stochastic gradients

Stochastic Gradient Descent in non-parametric problems

If we define ℓ by

$$\ell(f) = \int (y - f(x))^2 d\mathbb{P}(y, x)$$

Then, being in an RKHS gives a simple gradient¹

$$\nabla \ell(f) = -2 \int (y - f(x)) K(x, \cdot) d\mathbb{P}(y, x)$$

Q1: What to do if not in RKHS?

¹the gradient is taken wrt the RKHS inner-prod, $\langle \cdot, \cdot \rangle_{\mathcal{F}}$

Stochastic Gradient Descent in non-parametric problems

Ok, so gradient descent would look like:

To find $f^* = \operatorname{argmin}_{f \in \mathcal{F}} E [(y - f(x))^2]$

1. Initialize f_0
2. For $k = 1, \dots$, update by

$$f_k = f_{k-1} + t_k \left(\int (y - f_{k-1}(x)) K(x, \cdot) d\mathbb{P}(y, x) \right)$$

for some step-size t_k

But we don't have access to $d\mathbb{P}(y, x)!!$

Stochastic Gradient Descent in non-parametric problems

Instead we use a stochastic gradient (calculated from data):

$$E [(y_i - f(x_i))K(x_i, \cdot)] = \int (y - f_{k-1}(x))K(x, \cdot)d\mathbb{P}(y, x)$$

This is where the “data” come in!

Stochastic Gradient Descent in non-parametric problems

Ok, so **stochastic**-gradient descent would look like:

To find $f^* = \operatorname{argmin}_{f \in \mathcal{F}} E [(y - f(x))^2]$

1. Initialize f_0
2. For $k = 1, \dots$, update by

$$f_k = f_{k-1} + t_k(y_i - f_{k-1}(x_i))K(x_i, \cdot)$$

for some step-size t_k

Stochastic Gradient Descent in non-parametric problems

Authors ² have painstakingly shown that...

with some care...³

This converges at minimax optimal rate*

²*Nonparametric stochastic approximation with large step-sizes* by Francis Bach's group is probably most comprehensive

³generally with “averaging”

Stochastic Gradient Descent in non-parametric problems

Authors ² have painstakingly shown that...

with some care...³

This converges at minimax optimal rate*

However! This uses n basis vectors after n steps

So requires $\sim n^2$ computation!

That improves over n^3 , but far from $n^{\frac{2\alpha+2}{2\alpha+1}}$

²*Nonparametric stochastic approximation with large step-sizes* by Francis Bach's group is probably most comprehensive

³generally with “averaging”

Stochastic Gradient Descent in non-parametric problems

Can we more cleverly choose our stochastic gradient direction?

In particular, using spectrum of an operator involving $K(\cdot, \cdot)$?

Perhaps use same ψ_k as in the original projection estimator
and just add basis elements as we go.

Problem: These basis elements are not orthogonal wrt $dP(x)$

As we add new elements, old weights may/will change

Stochastic Gradient Descent in non-parametric problems

Can we more cleverly choose our stochastic gradient direction?

In particular, using spectrum of an operator involving $K(\cdot, \cdot)$?

We were using eigenstructure of operator

$$T : Tf \equiv \int_X f(x)K(x, \cdot)dx$$

Why not switch to

$$\tilde{T} : \tilde{T}f \equiv \int_X f(x)K(x, \cdot)dP(x)$$

Problem: Don't have $dP(x)$!

Though this would be the right “diagonalization”

Noah's Approach

Can we more cleverly choose our stochastic gradient direction?

In particular, using spectrum of an operator involving $K(\cdot, \cdot)$?

How about we use a “doubly stochastic algorithm” ...

Sample ψ_j to use at each step, while maintaining $E[\tilde{\nabla}(f)] = \nabla(f)$

Noah's Approach

Can we more cleverly choose our stochastic gradient direction?

In particular, using spectrum of an operator involving $K(\cdot, \cdot)$?

How about we use a “doubly stochastic algorithm” ...

Could use

$$\tilde{\nabla}(f) = (y_i - f(x_i)) \sum_{j=1}^{\infty} z_j \psi_j(x_i)$$

with z_j indep and $E[z_j] = \lambda_j$ (probably scaled bernoulli)

Hopefully need to store fewer than n basis vectors

(in expectation...)

More Questions

There is still an open question in comparing the spectrum of

$$\tilde{T} : \tilde{T}f \equiv \int_X f(x)K(x, \cdot) dP(x)$$

to the spectrum of

$$T : Tf \equiv \int_X f(x)K(x, \cdot) dx$$

Previous work requires assumptions on spectrum of both...

Can we make assumption on T , and thus bound spectrum of $\tilde{T}$?

More Questions

Also, what if $P(f)$ that is not embeddable in an RKHS?

eg. how about $\int_X |f'(x)| dx$?

We know for these problems that information is compressible...

But! dangerous to apriori specify basis elements

I suspect that “optimal” time/space complexity will involve adaptive choice of basis elements