Low-rank and sparse methods for high-dimensional approximation and model order reduction

Lecture 2

Sparse approximation

Sparse approximation

We want to approximate a function u in a certain function (or vector) space X equipped with a norm $\|\cdot\|$.

Consider a set $\mathcal{D}$ of functions in X (called a dictionary) such that the linear span of $\mathcal{D}$ is dense in X.

As a basic example, consider the case where $\mathcal{D} = \{\psi_k\}_{k\geq 0}$ is a basis of X.

Sparse approximation methods rely on the fact that a good approximation (or even an exact decomposition) of the solution can be obtained by only considering a small subset of functions in the dictionary:

$$u \approx u_n = \sum_{\psi\in\mathcal{D}_n} c_\psi \psi; \quad \mathcal{D}_n \subset \mathcal{D}, \quad \#\mathcal{D}_n = n.$$

u_n is called a n-term approximation of u. We say that u_n is n-sparse relatively to $\mathcal{D}$.

Dictionaries for high-dimensional approximation

For high-dimensional approximation problems, dictionaries must have low-dimensional parametrizations.

Typical choices are:

- Tensorized basis (e.g. polynomial basis, wavelets basis, ...):

$$\mathcal{D} = \{\psi_\alpha(x) = \psi_{\alpha_1}(x_1) \dots \psi_{\alpha_d}(x_d) : \alpha \in \mathcal{F}\}$$

- Separated (rank-one) functions:

$$\mathcal{D} = \{u_1(x_1) \dots u_d(x_d) : u_1 \in \mathcal{H}_1, \dots, u_d \in \mathcal{H}_d\}$$

- Perceptrons (for neural networks):

$$\mathcal{D} = \{\sigma(a^T x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}\}$$

- Functions of linear combinations of variables (for projection pursuit):

$$\mathcal{D} = \{g(a^T x) : a \in \mathbb{R}^d, g \in \mathcal{H}\}$$

- ...

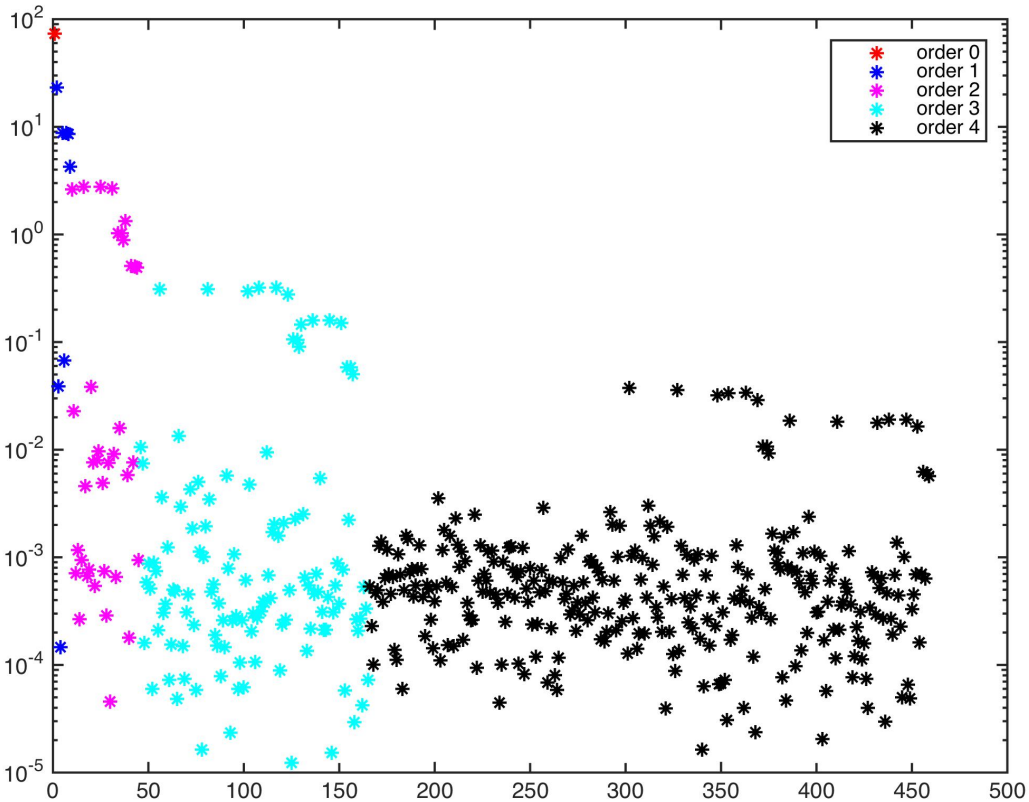
A motivating example

We consider the Borehole function which models the water flow through a borehole:

$$u(X) = \frac{2\pi T_u(H_u - H_l)}{\ln(r/r_w)\left(1 + \frac{2LT_u}{\ln(r/r_w)r_w^2 K_w} + \frac{T_u}{T_l}\right)}, \quad X = (r_w, r, , T_u, H_u, T_l, H_l, L, K_w) \sim P_X$$

Consider $\mathcal{D} = \{\psi_k\}$ as the set of multivariate polynomials (orthogonal w.r.t P_X).

The following plot shows the coefficients c_k of u associated with polynomial functions ψ_k of total degree less than 4 (colors indicate the total degree of ψ_k).



Outline

Outline

Convergence of best n-term approximation

For a given dictionary $\mathcal{D}$, the ideal performance of sparse approximation methods is quantified by the best n-term approximation error

$$\sigma_n(u) = \min_{v \in \Sigma_n} \|u - v\|$$

where $\Sigma_n = \left\{ \sum_{\psi \in \mathcal{D}_n} c_\psi \psi : \#\mathcal{D}_n = n \right\}$ is the set of n-sparse elements.

Let us assume that u admits the decomposition

$$u = \sum_{k=1}^{\infty} c_k \psi_k, \quad \psi_k \in \mathcal{D}.$$

Then

$$\sigma_n(u) = \min_{v \in \Sigma_n} \|u - v\| \leq \min_{\#\Lambda = n} \|u - u_\Lambda\|, \qquad \text{with } u_\Lambda = \sum_{k \in \Lambda} c_k \psi_k.$$

Assuming that the elements of $\mathcal{D}$ are normalized, we obtain

$$\sigma_n(u) \leq \min_{\#\Lambda = n} \| \sum_{k \notin \Lambda} c_k \psi_k \| \leq \min_{\#\Lambda = n} \sum_{k \notin \Lambda} |c_k| = \sum_{k \notin \Lambda_n} |c_k|,$$

where Λ_n corresponds to the n largest coefficients $|c_k|$.

Convergence of best n-term approximation

Equivalently, we obtain

$$\sigma_n(u) \leq \sum_{k=n+1}^{\infty} c_k^*$$

where $\mathbf{c}^* = (c_k^*)_{k\geq 1}$ is a decreasing rearrangement of $(|c_k|)_{k\geq 1}$.

Therefore, if u admits a decomposition with rapidly decaying coefficients, we can expect a fast convergence of best n-term approximation error.

In particular:

- If c_n^* decays exponentially, $\sigma_n(u)$ decays exponentially with the same rate.
- If $\mathbf{c}^* \in \ell_p$ with $0 < p < 1$, then $\sigma_n(u) \leq \|\mathbf{c}^*\|_{\ell_p} n^{-r}$ with $r = 1/p - 1$.

Of course, for a given function, the performance of sparse approximation methods strongly depends on the choice of the dictionary...

Quasi best n-term approximation

In practice, n-term approximations u_n are defined by

$$\mathcal{J}(u_n) = \min_{v \in \Sigma_n} \mathcal{J}(v) \tag{1}$$

where $\mathcal{J}$ is a computable functional.

If $\mathcal{J}(v)$ measures a distance from v to u such that

$$\alpha\|u - v\| \leq \mathcal{J}(v) \leq \beta\|u - v\|, \tag{2}$$

then the solution u_n of (1) is such that

$$\|u - u_n\| \leq \frac{1}{\alpha}\mathcal{J}(u_n) = \frac{1}{\alpha} \min_{v \in \Sigma_n} \mathcal{J}(v) \leq \frac{\beta}{\alpha} \min_{v \in \Sigma_n} \|u - v\|,$$

which means that u_n is a quasi-optimal n-term approximation.

Property (2) is satisfied when using variational methods for solving operator equations $Au = b$, where

$$\mathcal{J}(v) = \|b - Av\|$$

with A an operator such that

$$\alpha\|v\| \leq \|Av\| \leq \beta\|v\|.$$

Approximation using partial information

Sometimes, we only have partial information on the function u, such as the evaluations $y^k = u(x^k)$ of the function u at some points x^k, $k = 1, \ldots, m$.

More generally, we assume that m measurements of u are given by

$$y = Au$$

where $A : X \to \mathbb{R}^m$ is a linear operator.

The functional

$$\mathcal{J}(v) = \|y - Av\|^2 = \frac{1}{m} \sum_{k=1}^{m} (y^k - (Av)_k)^2$$

then provides a distance between the measurements y and the prediction Av. When too few observations are available, the optimization problem

$$\min_{v \in X} \mathcal{J}(v) \tag{3}$$

is ill-posed.

Example 1 (Least-Squares)

In the case where $(Av) = v(x^k)$ is the evaluation of v at point x^k, then (3) is a standard least-squares problem.

Approximation using partial information

Imposing the approximation to be n-sparse by solving

$$\min_{v \in \Sigma_n} \|y - Av\|^2 \tag{4}$$

is a possible way to make the problem well-posed.

Assuming that u is r-sparse and that A satisfies

$$(1-\delta)\|v\|^2 \leq \|Av\|^2 \leq (1+\delta)\|v\|^2 \quad \text{for all } v \in \Sigma_s, \tag{5}$$

which is a restricted isometry property, then for all $n \leq s - r$, problem (4) admits a solution u_n such that

$$\|u - u_n\| \leq C \min_{v \in \Sigma_n} \|u - v\|,$$

with $C^2 = \frac{1+\delta}{1-\delta}$.

Property (5) depends on the dictionary and of the measurement operator A.

Approximation using partial information

Example 2 (Least-Squares with orthonormal basis)

Consider the case where $\mathcal{D} = \{\psi_i\}_{i=1}^N$ is an orthonormal basis in $L^2_{P_X}(\mathcal{X})$. A function $v = \sum_{i=1}^N a_i \psi_i$ is such that $\|v\| = \|\mathbf{a}\|_2$, with $\mathbf{a} \in \mathbb{R}^N$.

Assume that the operator A provides evaluations at points $\{x^k\}_{k=1}^m$, i.e. $Av = (v(x^k))_{k=1}^m$. Then

$$\|Av\|^2 - \|v\|^2 = \frac{1}{m}\|\mathbf{\Phi a}\|_2^2 - \|\mathbf{a}\|_2^2 = ((\mathbf{G} - \mathbf{I})\mathbf{a}, \mathbf{a}),$$

where $\mathbf{\Phi} = (\psi_i(x^k)) \in \mathbb{R}^{m \times N}$ is the matrix of evaluations of functions ψ_i at points x^k, and

$$\mathbf{G} = \frac{1}{m}\mathbf{\Phi}^T\mathbf{\Phi} = (\frac{1}{m}\sum_{k=1}^m \psi_i(x^k)\psi_j(x^k))_{ij}.$$

The problem is then to analyze how far the restriction of the matrix $\mathbf{G}$ to the subset of sparse vectors is from the identity matrix $\mathbf{I}$.

When x^k are m samples of $X \sim P_X$, $\mathbf{G}$ is an unbiased and convergent estimate of $\mathbf{I}$. Under some assumptions and results from random matrix theory, restricted isometry property for s-sparse elements can be proved to be satisfied with high probability for a number of measurements m in $O(s)$.

Statistical point of view

We consider a pair of random variables (X, Y) with values in $(\mathcal{X}, \mathbb{R})$ such that

$$Y = u(X) + \epsilon$$

where ϵ represents a noise.

The aim is to estimate (or learn) u from a sample $S = \{(x^1, y^1), \ldots, (x^n, y^n)\}$ of (X, Y) (a training set).

For that, we minimize an empirical risk

$$\mathcal{J}(v) := \widehat{\mathcal{R}}_n(v) = \frac{1}{n} \sum_{k=1}^{n} \ell(y^k, v(x^k))$$

where $\ell(y, v(x))$ is a certain loss function which measures a certain error (a cost) when replacing y by the prediction $v(x)$.

The empirical risk is a statistical estimate of the risk functional

$$\mathcal{R}(v) = \mathbb{E}(\ell(Y, v(X)))$$

Statistical point of view

For least-square regression, we consider the loss $\ell(y, v(x)) = (y - v(x))^2$, so that

$$\mathcal{J}(v) = \frac{1}{n}\sum_{k=1}^{n}(y^k - v(x^k))^2$$

and

$$\mathcal{R}(v) = \mathbb{E}((Y - v(X))^2).$$

Assuming that ϵ is zero mean and independent of X, we have

$$\mathcal{R}(v) = \mathbb{E}((u(X) - v(X))^2) + Var(\epsilon),$$

and the empirical risk minimization is a statistical approach for L^2 approximation.

About solving the best n-term approximation problem

Assuming that $\mathcal{D} = \{\psi_k\}_{k=1}^{N}$, solving the best n-term approximation problem

$$\min_{v \in \Sigma_n} \mathcal{J}(v) \tag{6}$$

a priori requires testing all possible subsets of n functions in $\mathcal{D}$. When $N < \infty$, that means $\binom{N}{n}$ possibilities (*NP*-hard problem). And obviously, the situation is even worse when N is infinite or $\mathcal{D}$ is not a countable set !

In practice, we rely on algorithms which produce approximate solutions to problem (6), such as

- Greedy algorithms,
- Convex relaxation methods,
- Working set algorithms.

Outline

Greedy algorithms

Given a dictionary $\mathcal{D}$ in X, greedy algorithms aim to build a sequence of suboptimal yet good n-terms approximations $(u_n)_{n\geq 0}$ with

$$u_n \in X_n := span\{\psi_1, \ldots, \psi_n\}$$

where the elements $(\psi_n)_{n\geq 1}$ are selected one-by-one in $\mathcal{D}$.

In the case where $\mathcal{D}$ is finite, greedy algorithms allow to break the combinatorial complexity of the best n-term approximation problem. When $\mathcal{D}$ is not a finite set, it provides a simple way to construct n-term approximations.

There are several variants of greedy algorithms depending on how to select the ψ_n and how to compute the approximation in X_n.

Standard greedy algorithms

A pure greedy algorithm (PGA) defines

$$u_n = u_{n-1} + c_n \psi_n$$

with

$$\mathcal{J}(u_{n-1} + c_n \psi_n) = \min_{\psi \in \mathcal{D}, c \in \mathbb{R}} \mathcal{J}(u_{n-1} + c\psi). \tag{7}$$

The orthogonal greedy algorithm (OGA) selects ψ_n based on (7) but the n-term approximation u_n is defined as the projection onto the generated subspace $X_n = span\{\psi_1, \ldots, \psi_n\}$

$$\mathcal{J}(u_n) = \arg \min_{v \in X_n} \mathcal{J}(v).$$

When X is a Hilbert space, $\mathcal{J}(v) = \|u - v\|$ and $\mathcal{D}$ is an orthonormal basis of X, greedy algorithms provide a sequence of best n-term approximations u_n. For other dictionaries $\mathcal{D}$, the obtained n-term approximations may be far from best n-term approximations.

Outline

Reformulations of best n-term approximation problem

We consider the case of a finite dictionary $\mathcal{D} = \{\psi_k\}_{k=1}^N$.

We denote by $\Psi : \mathbb{R}^N \to X$ the operator which associates to a set of coefficients $\mathbf{a} = (a_k)_{k=1}^N$ the element

$$\Psi\mathbf{a} = \sum_{k=1}^{N} a_k \psi_k \in X.$$

The set of n-sparse elements is

$$\Sigma_n = \{\Psi\mathbf{a} : \|\mathbf{a}\|_0 \leq n\}$$

where $\|\cdot\|_0$ is the so-called "ℓ_0-norm" of the set of coefficients

$$\|\mathbf{a}\|_0 = \#\{k : a_k \neq 0\}.$$

The best n-term approximation problem

$$\min_{v \in \Sigma_n} \mathcal{J}(v)$$

is then equivalent to

$$\min_{\mathbf{a}} \mathcal{J}(\Psi\mathbf{a}) \quad \text{subject to} \quad \|\mathbf{a}\|_0 \leq n.$$

Reformulations of best n-term approximation problem

A related formulation if given by the unconstrained minimization problem

$$\min_{\mathbf{a}\in\mathbb{R}^m} \mathcal{J}(\Psi\mathbf{a}) + \lambda\|\mathbf{a}\|_0. \tag{8}$$

When increasing λ, problem (8) provides sparser and sparser solutions $\mathbf{a}$.

If X is a Hilbert space, $\mathcal{J}(v) = \|u - v\|^2$, with $\|\cdot\|$ associated with an inner product $(\cdot,\cdot)$, and $\mathcal{D}$ is an orthonormal basis, then the solution of (8) is

$$a_i = HT_{\sqrt{\lambda}}(c_i), \quad c_i = (u, \psi_i),$$

where

$$HT_\tau(t) = t\, 1_{|t|>\tau}$$

is the hard thresholding function, which means

$$a_i = \begin{cases} c_i & \text{if } |c_i| > \sqrt{\lambda} \\ 0 & \text{if } |c_i| \leq \sqrt{\lambda} \end{cases}$$

Convex relaxation

The problem can be replaced by

$$\min_{\mathbf{a}\in\mathbb{R}^m} \mathcal{J}(\Psi\mathbf{a}) + \lambda\|\mathbf{a}\|_1. \tag{9}$$

If $\mathcal{J}$ is convex, it is a convex optimization problem.

If X is a Hilbert space, $\mathcal{J}(v) = \frac{1}{2}\|u - v\|^2$, with $\|\cdot\|$ associated with an inner product $(\cdot,\cdot)$, and $\mathcal{D}$ is an orthonormal basis, then the solution of (9) is

$$a_i = ST_\lambda(c_i), \quad c_i = (u, \psi_i),$$

where

$$ST_\lambda(t) = (|t| - \tau)_+ \operatorname{sign}(t)$$

is the soft thresholding function, which means

$$a_i = \begin{cases} c_i - \lambda & \text{if } c_i > \lambda \\ 0 & \text{if } |c_i| \leq \lambda \\ c_i + \lambda & \text{if } c_i < -\lambda \end{cases}$$

Increasing λ yields sparser and sparser solutions.

Convex relaxation

About algorithms for solving problem (9):

- It is a non-differentiable optimization problem.
- For a convex functional $\mathcal{J}$, algorithms for non-differentiable convex optimization are available (e.g. proximal methods).
- In the case where $\mathcal{J}(\Psi\mathbf{a}) = \|\mathbf{y} - \mathbf{\Phi a}\|_2^2$, (9) is the LASSO problem. The LARS homotopy algorithm provides the set of solutions for all values of λ.

About the selection of regularization parameter:

- Computing the solution for many values of λ provides a set of solutions $\mathbf{a}^\lambda$ with different sparsity patterns $\Lambda^\lambda = \{k : a_k^\lambda \neq 0\}$.
- A particular solution can be selected using error estimates.
- For a given pattern Λ^λ, the best approximation can be computed by solving

$$\min_{\mathbf{a}} \mathcal{J}(\Psi\mathbf{a}) \quad \text{subject to } a_k = 0 \text{ for } k \notin \Lambda^\lambda \tag{10}$$

- In a statistical framework, validation or cross-validation error estimates can be used. Note that for usual functionals $\mathcal{J}$, cross-validation error estimates can be obtained very efficiently for the solutions of problem (10).

Convex relaxation

Other notions of sparsity can be imposed by considering problems of the form

$$\min_{\mathbf{a}\in\mathbb{R}^N} \mathcal{J}(\Psi\mathbf{a}) + \lambda\Omega(\mathbf{a}) \tag{11}$$

with a suitable choice for Ω.

- Weighted sparsity with weighted ℓ_1 norm:

$$\Omega(\mathbf{a}) = \|\mathbf{a}\|_{1,\omega} = \sum_{k=1}^{N} \omega_k |a_k|$$

- Group sparsity with $\ell_1 - \ell_2$ norms:

$$\Omega(\mathbf{a}) = \sum_{\nu=1}^{K} \|\mathbf{a}_{J_\nu}\|_2, \quad \bigcup_{\nu=1}^{K} J_\nu = \{1, \ldots, N\}$$

- ...

Outline

Working set algorithms

For a finite (or countable) dictionary $\mathcal{D} = \{\psi_k\}_{k \geq 1}$, working set algorithms are algorithms which construct an increasing sequence of index sets $(\Lambda_n)_{n \geq 1}$.

For a given pattern Λ, an approximation $u_\Lambda = \sum_{k \in \Lambda} a_k \psi_k$ is computed using interpolation, regression or other projection methods.

At step n, Λ_n is defined by

$$\Lambda_n = \Lambda_{n-1} \cup A_n$$

where A_n is a set of new indices picked in a set of candidate indices N_n, based on some selection criterium.

If for each $k \in N_n$ we can estimate a profit $e(k)$ of adding k to Λ_n, we can choose $A_n = \{k^n\}$ with

$$e(k^n) = \max_{k \in N_n} e(k),$$

or even A_n as the set of all indices $k \in N_n$ such that

$$e(k) \geq \theta \max_{j \in N_n} e(j).$$