Lecture 3: Discrete Compressed Sensing

Ronald DeVore

Traditional Paradigm for Signal Processing

- Model signals as bandlimited functions $x(t)$
- Support of $\hat{x}$ contained in $[-\Omega\pi, \Omega\pi]$
- Shannon-Nyquist
- Uniform time samples with spacing $h \leq \frac{1}{\Omega}$ allows for exact reconstruction
- A/D Converters: sample and quantize
- Problem: If Ω is too large we cant build circuitry to sample faithfully at the desired rate

Compressed Sensing

- Compressed Sensing seeks a way out of this dilemma
- Built on two new ingredients
- New model classes for signals
- Signals are sparse in some representation system: basis (frame)
- New meaning of samples
- Sample is a linear functional applied to the signal

Discrete Sensing

- $x \in \mathbb{R}^N$ with N large

Discrete Sensing

- $x \in \mathbb{R}^N$ with N large
- We are able to ask n non-adaptive questions about x

Discrete Sensing

- $x \in \mathbb{R}^N$ with N large
- We are able to ask n non-adaptive questions about x
- Question means inner product $v \cdot x$ with $v \in \mathbb{R}^N$ - called sample

Discrete Sensing

- $x \in I\!R^N$ with N large
- We are able to ask n non-adaptive questions about x
- Question means inner product $v \cdot x$ with $v \in I\!R^N$ - called sample
- Any such sampling is given by is an $n \times N$ matrix Φ : the entries in $y = \Phi x$ are the answers to our questions

Discrete Sensing

- $x \in \mathbb{R}^N$ with N large
- We are able to ask n non-adaptive questions about x
- Question means inner product $v \cdot x$ with $v \in \mathbb{R}^N$ - called sample
- Any such sampling is given by is an $n \times N$ matrix Φ : the entries in $y = \Phi x$ are the answers to our questions
- We are interested in the good / best matrices Φ, i.e. what are the best questions to ask??

Discrete Sensing

- $x \in \mathbb{R}^N$ with N large
- We are able to ask n non-adaptive questions about x
- Question means inner product $v \cdot x$ with $v \in \mathbb{R}^N$ - called sample
- Any such sampling is given by is an $n \times N$ matrix Φ : the entries in $y = \Phi x$ are the answers to our questions
- We are interested in the good / best matrices Φ, i.e. what are the best questions to ask??
- Here good roughly means that the samples $y = \Phi x$ contain enough information to approximate x well

Discrete Sensing

- $x \in \mathbb{R}^N$ with N large
- We are able to ask n non-adaptive questions about x
- Question means inner product $v \cdot x$ with $v \in \mathbb{R}^N$ - called sample
- Any such sampling is given by is an $n \times N$ matrix Φ : the entries in $y = \Phi x$ are the answers to our questions
- We are interested in the good / best matrices Φ, i.e. what are the best questions to ask??
- Here good roughly means that the samples $y = \Phi x$ contain enough information to approximate x well
- Two issues: (i) Enough information in y; (ii) How to extract this information: Decoder

What is going on?

- Since $\Phi : \mathbb{R}^N \to \mathbb{R}^n$ many x give the same measurements y

What is going on?

- Since $\Phi : I\!R^N \to I\!R^n$ many x give the same measurements y
- $\mathcal{N} := \{\eta : \Phi\eta = 0\}$ the null space of Φ $\dim(\mathcal{N}) \geq N - n$

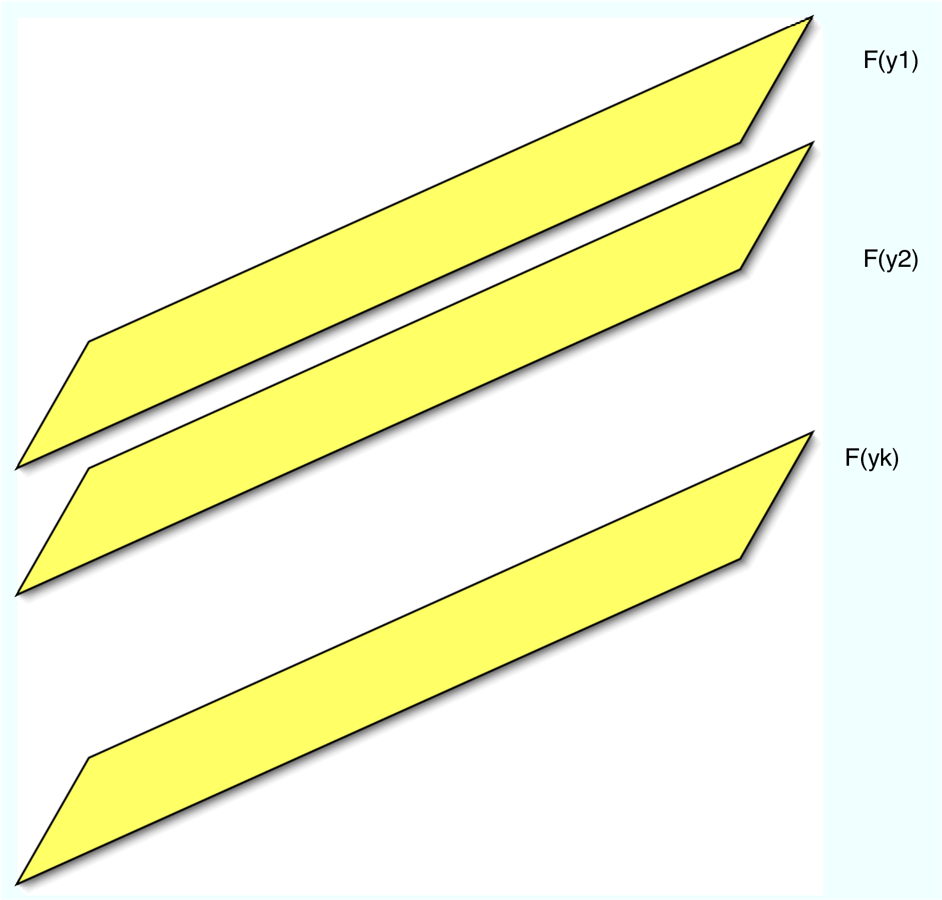
What is going on?

- Since $\Phi : I\!R^N \to I\!R^n$ many x give the same measurements y
- $\mathcal{N} := \{\eta : \Phi\eta = 0\}$ the null space of Φ $\dim(\mathcal{N}) \geq N - n$
- $\mathcal{F}(y) := \{x : \Phi x = y\} = x_0 + \mathcal{N}$ for any $x_0 \in \mathcal{F}(y)$

What is going on?

- Since $\Phi : I\!R^N \to I\!R^n$ many x give the same measurements y
- $\mathcal{N} := \{\eta : \Phi\eta = 0\}$ the null space of Φ $\dim(\mathcal{N}) \geq N - n$
- $\mathcal{F}(y) := \{x : \Phi x = y\} = x_0 + \mathcal{N}$ for any $x_0 \in \mathcal{F}(y)$
- The hyperplanes $\mathcal{F}(y)$ with $y \in I\!R^n$ stratify $I\!R^N$

The sets $\mathcal{F}(y)$



What is going on?

- Since $\Phi : I\!R^N \to I\!R^n$ many x are encoded with same y
- $\mathcal{N} := \{\eta : \Phi\eta = 0\}$ the null space of Φ
- $\mathcal{F}(y) := \{x : \Phi x = y\} = x_0 + \mathcal{N}$ for any $x_0 \in \mathcal{F}(y)$
- The hyperplanes $\mathcal{F}(y)$ with $y \in I\!R^n$ stratify $I\!R^N$
- Decoder is any (possibly nonlinear) mapping Δ from $I\!R^n \to I\!R^N$

What is going on?

- Since $\Phi : I\!R^N \to I\!R^n$ many x are encoded with same y
- $\mathcal{N} := \{\eta : \Phi\eta = 0\}$ the null space of Φ
- $\mathcal{F}(y) := \{x : \Phi x = y\} = x_0 + \mathcal{N}$ for any $x_0 \in \mathcal{F}(y)$
- The hyperplanes $\mathcal{F}(y)$ with $y \in I\!R^n$ stratify $I\!R^N$
- Decoder is any (possibly nonlinear) mapping Δ from $I\!R^n \to I\!R^N$
- $\bar{x} := \Delta(\Phi(x))$ is our approximation to x from the information extracted

What is going on?

- Since $\Phi : I\!R^N \to I\!R^n$ many x are encoded with same y
- $\mathcal{N} := \{\eta : \Phi\eta = 0\}$ the null space of Φ
- $\mathcal{F}(y) := \{x : \Phi x = y\} = x_0 + \mathcal{N}$ for any $x_0 \in \mathcal{F}(y)$
- The hyperplanes $\mathcal{F}(y)$ with $y \in I\!R^n$ stratify $I\!R^N$
- Decoder is any (possibly nonlinear) mapping Δ from $I\!R^n \to I\!R^N$
- $\bar{x} := \Delta(\Phi(x))$ is our approximation to x from the information extracted
- Pessimism: all $x \in \mathcal{F}(y)$ are approximated by the same $\bar{x}$

Structure

- With no additional information about x it is doubtful we can say anything

Structure

- With no additional information about x it is doubtful we can say anything
- Fortunately the x we are interested in have structure

Structure

- With no additional information about x it is doubtful we can say anything
- Fortunately the x we are interested in have structure
- Typically x can be well represented by sparse linear combination of certain building blocks - for our purposes these building blocks are a basis

Structure

- With no additional information about x it is doubtful we can say anything
- Fortunately the x we are interested in have structure
- Typically x can be well represented by sparse linear combination of certain building blocks - for our purposes these building blocks are a basis
- In some(many) problems we do not necessarily know the right basis

Structure

- With no additional information about x it is doubtful we can say anything
- Fortunately the x we are interested in have structure
- Typically x can be well represented by sparse linear combination of certain building blocks - for our purposes these building blocks are a basis
- In some(many) problems we do not necessarily know the right basis
- For the most part we shall assume the basis is known to us

First Measure of Optimality

- To begin with we shall assume x is sparse with respect to the canonical basis on $\mathbb{R}^N$

First Measure of Optimality

- To begin with we shall assume x is sparse with respect to the canonical basis on $\mathbb{R}^N$
- Any other basis could be handled by transformation (if the basis is known)

First Measure of Optimality

- To begin with we shall assume x is sparse with respect to the canonical basis on $\mathbb{R}^N$
- Any other basis could be handled by transformation (if the basis is known)
- The support of x is $\mathrm{supp(x) := \{i : \ x_i \neq 0\}}$

First Measure of Optimality

- To begin with we shall assume x is sparse with respect to the canonical basis on $\mathbb{R}^N$
- Any other basis could be handled by transformation (if the basis is known)
- The support of x is $\mathrm{supp(x) := \{i : x_i \neq 0\}}$
- $\Sigma_k := \{x : \mathrm{\#supp(x) \leq k}\}$

First Measure of Optimality

- To begin with we shall assume x is sparse with respect to the canonical basis on $\mathbb{R}^N$
- Any other basis could be handled by transformation (if the basis is known)
- The support of x is $\mathrm{supp(x) := \{i : \ x_i \neq 0\}}$
- $\Sigma_k := \{x : \ \#\mathrm{supp(x)} \leq \mathrm{k}\}$
- Note that Σ_k is a union of k dimensional subspaces: $\Sigma_k = \cup_{\#(T)=k} X_T$ where $X_T = \{x : \mathrm{supp(x)} \subset \mathrm{T}\}$

First Question

- First Question: Given k, N for which Φ does the information $y = \Phi(x)$ uniquely determine x for all $x \in \Sigma_k$

First Question

- First Question: Given k, N for which Φ does the information $y = \Phi(x)$ uniquely determine x for all $x \in \Sigma_k$
- Note that this is equivalent to $\mathcal{F}(y)$ contains at most one vector from Σ_k for all $y \in I\!R^n$

The Answer

- $\Phi = [v_1, \ldots, v_N], \quad v_1, \ldots, v_N$ columns of Φ

The Answer

- $\Phi = [v_1, \dots, v_N], \quad v_1, \dots, v_N$ columns of Φ
- If T is a set of column indices

The Answer

- $\Phi = [v_1, \ldots, v_N], \quad v_1, \ldots, v_N$ columns of Φ
- If T is a set of column indices
- $\Phi_T = [v_{i_1}, \ldots, v_{i_m}]$ is the $n \times \#(T)$ submatrix of Φ formed from the columns with index $T = \{i_1, \ldots, i_m\}$

The Answer

- $\Phi = [v_1, \ldots, v_N], \quad v_1, \ldots, v_N$ columns of Φ
- If T is a set of column indices
- $\Phi_T = [v_{i_1}, \ldots, v_{i_m}]$ is the $n \times \#(T)$ submatrix of Φ formed from the columns with index $T = \{i_1, \ldots, i_m\}$
- $\Phi_T^* \Phi_T := (\langle v_i, v_j \rangle)_{i,j \in T}$

The Answer

- $\Phi = [v_1, \dots, v_n], \quad v_1, \dots, v_N$ columns of Φ
- If T is a set of column indices
- $\Phi_T = [v_{i_1}, \dots, v_{i_m}]$ is the $\#(T) \times n$ submatrix of Φ formed from the columns with index $T = \{i_1, \dots, i_m\}$
- $\Phi_T^* \Phi_T := (\langle v_i, v_j \rangle)_{i,j \in T}$ of size $\#(T) \times \#(T)$

Theorem: If Φ is any $n \times N$ matrix, then the following are equivalent:

(i) $\mathcal{F}(y)$ contains at most one element of Σ_k for all $y \in I\!R^n$
(ii) $\Sigma_{2k} \cap \mathcal{N}(\Phi) = \{0\}$,
(iii) For any set T with $\#T = 2k$, the matrix Φ_T has rank $2k$.
(iv) For any set T with $\#T = 2k$, the $2k \times 2k$ matrix $\Phi_T^* \Phi_T$ is invertible: all its eigenvalues are positive.

The Proof

- The equivalence of (ii-iv) is linear algebra

The Proof

- The equivalence of (ii-iv) is linear algebra
- (i) $\rightarrow$ (ii): Assume (i). To prove (ii), let $\eta \in \mathcal{N} \cap \Sigma_{2k}$. We can write $\eta = \eta_0 - \eta_1$ with $\eta_0, \eta_1 \in \Sigma_k$ and $\Phi(\eta_0) = \Phi(\eta_1)$ Hence by (i), $\eta_0 = \eta_1$ and therefore $\eta = 0$

The Proof

- The equivalence of (ii-iv) is linear algebra
- (i) $\to$ (ii): Assume (i). To prove (ii), let $\eta \in \mathcal{N} \cap \Sigma_{2k}$. We can write $\eta = \eta_0 - \eta_1$ with $\eta_0, \eta_1 \in \Sigma_k$ and $\Phi(\eta_0) = \Phi(\eta_1)$ Hence by (i), $\eta_0 = \eta_1$ and therefore $\eta = 0$
- (ii) $\to$ (i): Assume (ii) and let $y \in \mathbb{R}^n$. Suppose that there are two vectors $x_1, x_2 \in \mathbb{R}^N$ with $\Phi(x_1) = \Phi(x_2) = y$. Then $\eta := x_1 - x_2 \in \mathcal{N}$ and by (ii) $\eta = 0$. Thus $x_1 = x_2$

Optimal Matrices

- Given k, we are interested in the smallest n for which we can construct matrices with properties (i)-(iv)

Optimal Matrices

- Given k, we are interested in the smallest n for which we can construct matrices with properties (i)-(iv)
- Given k we construct matrices Φ of size $2k \times N$ with the properties of the theorem

Optimal Matrices

- Given k, we are interested in the smallest n for which we can construct matrices with properties (i)-(iv)
- Given k we construct matrices Φ of size $2k \times N$ with the properties of the theorem
- We need N vectors in $\mathbb{R}^{2k}$ such that any $2k$ of them are linearly independent

Optimal Matrices

- Given k, we are interested in the smallest n for which we can construct matrices with properties (i)-(iv)
- Given k we construct matrices Φ of size $2k \times N$ with the properties of the theorem
- We need N vectors in $\mathbb{R}^{2k}$ such that any $2k$ of them are linearly independent
- Vandermonde matrix. Choose $x_1 < x_2 < \cdots < x_N$

$$\Phi := \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_1^{2k-1} & x_2^{2k-1} & \cdots & x_N^{2k-1} \end{pmatrix}$$

Optimal Matrices

- Given k, we are interested in the smallest n for which we can construct matrices with properties (i)-(iv)
- Given k we construct matrices Φ of size $2k \times N$ with the properties of the theorem
- We need N vectors in $\mathbb{R}^{2k}$ such that any $2k$ of them are linearly independent
- Vandermonde matrix. Choose $x_1 < x_2 < \cdots < x_N$

$$\Phi := \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_1^{2k-1} & x_2^{2k-1} & \cdots & x_N^{2k-1} \end{pmatrix}$$

Naive Decoding

$$\Delta(y) := \underset{z \in \Sigma_k}{\mathrm{Argmin}} \|y - \Phi(z)\|_{\ell_2^n}$$

- $X_T := \{z : \ \mathrm{supp(z)} \subset \mathrm{T}\}$

Naive Decoding

$$\Delta(y) := \operatorname*{Argmin}_{z \in \Sigma_k} \|y - \Phi(z)\|_{\ell_2^n}$$

- $X_T := \{z : \ \mathrm{supp(z)} \subset \mathrm{T}\}$
- $x_T := \operatorname*{Argmin}_{z \in X_T} \|y - \Phi z\|_{\ell_2^n} \rightarrow x_T = [\Phi_T^* \Phi_T]^{-1} \Phi_T y$

Naive Decoding

$$\Delta(y) := \operatorname*{Argmin}_{z \in \Sigma_k} \|y - \Phi(z)\|_{\ell_2^n}$$

- $X_T := \{z : \text{supp(z)} \subset \text{T}\}$
- $x_T := \operatorname*{Argmin}_{z \in X_T} \|y - \Phi z\|_{\ell_2^n} \rightarrow x_T = [\Phi_T^* \Phi_T]^{-1} \Phi_T y$
- $T^* := \text{Argmin}_{\#(T)=k} \|y - \Phi(x_T)\|_{\ell_2^n}$

Naive Decoding

$$\Delta(y) := \operatorname*{Argmin}_{z \in \Sigma_k} \|y - \Phi(z)\|_{\ell_2^n}$$

- $X_T := \{z : \text{ supp(z)} \subset \text{T}\}$
- $x_T := \operatorname*{Argmin}_{z \in X_T} \|y - \Phi z\|_{\ell_2^n} \rightarrow x_T = [\Phi_T^* \Phi_T]^{-1} \Phi_T y$
- $T^* := \text{Argmin}_{\#(T)=k} \|y - \Phi(x_T)\|_{\ell_2^n}$
- $\Delta(y) := x_{T^*}$

Naive Decoding

$$\Delta(y) := \operatorname*{Argmin}_{z\in\Sigma_k} \|y - \Phi(z)\|_{\ell_2^n}$$

- $X_T := \{z :\ \mathrm{supp(z)} \subset \mathrm{T}\}$
- $x_T := \operatorname*{Argmin}_{z\in X_T} \|y - \Phi z\|_{\ell_2^n} \rightarrow x_T = [\Phi_T^* \Phi_T]^{-1} \Phi_T y$
- $T^* := \mathrm{Argmin}_{\#(T)=k} \|y - \Phi(x_T)\|_{\ell_2^n}$
- $\Delta(y) := x_{T^*}$
- However one can improve the decoding to be more computationally reasonable

Naive Decoding

$$\Delta(y) := \operatorname*{Argmin}_{z \in \Sigma_k} \|y - \Phi(z)\|_{\ell_2^n}$$

- $X_T := \{z : \text{ supp(z)} \subset \text{T}\}$
- $x_T := \operatorname*{Argmin}_{z \in X_T} \|y - \Phi z\|_{\ell_2^n} \rightarrow x_T = [\Phi_T^* \Phi_T]^{-1} \Phi_T y$
- $T^* := \text{Argmin}_{\#(T)=k} \|y - \Phi(x_T)\|_{\ell_2^n}$
- $\Delta(y) := x_{T^*}$
- However one can improve the decoding to be more computationally reasonable
- For Fourier matrix, the computation can be done in $O(N + k^3)$ operations

Naive Decoding

$$\Delta(y) := \operatorname*{Argmin}_{z \in \Sigma_k} \|y - \Phi(z)\|_{\ell_2^n}$$

- $X_T := \{z : \ \mathrm{supp(z)} \subset \mathrm{T}\}$
- $x_T := \operatorname*{Argmin}_{z \in X_T} \|y - \Phi z\|_{\ell_2^n} \rightarrow x_T = [\Phi_T^* \Phi_T]^{-1} \Phi_T y$
- $T^* := \mathrm{Argmin}_{\#(T)=k} \|y - \Phi(x_T)\|_{\ell_2^n}$
- $\Delta(y) := x_{T^*}$
- However one can improve the decoding to be more computationally reasonable
- For Fourier matrix, the computation can be done in $O(N + k^3)$ operations
- However there is a stability problem caused by poor conditioning of $\Phi_T^* \Phi_T$

Model II: Compressible Signals

- The sparse signal classes do not represent real signals: signals will typically have all entries nonzero but most will be small

Model II: Compressible Signals

- The sparse signal classes do not represent real signals: signals will typically have all entries nonzero but most will be small
- A compressible signal x is one that can be approximated well by elements from Σ_k:
  $$\sigma_k(x)_X := \inf_{z \in \Sigma_k} \|x - z\|_X$$

Model II: Compressible Signals

- The sparse signal classes do not represent real signals: signals will typically have all entries nonzero but most will be small
- A compressible signal x is one that can be approximated well by elements from Σ_k:
  $\sigma_k(x)_X := \inf_{z \in \Sigma_k} \|x - z\|_X$
- to measure distortion we shall take $X = \ell_q^N$

Model II: Compressible Signals

- The sparse signal classes do not represent real signals: signals will typically have all entries nonzero but most will be small
- A compressible signal x is one that can be approximated well by elements from Σ_k:
  $\sigma_k(x)_X := \inf_{z \in \Sigma_k} \|x - z\|_X$
- to measure distortion we shall take $X = \ell_q^N$
- Recall our classes $\mathcal{A}^s(X)$, $s > 0$ consisting of signals for which $\sigma_k(x)_X \leq M k^{-s}$

Model II: Compressible Signals

- The sparse signal classes do not represent real signals: signals will typically have all entries nonzero but most will be small
- A compressible signal x is one that can be approximated well by elements from Σ_k:
  $\sigma_k(x)_X := \inf_{z \in \Sigma_k} \|x - z\|_X$
- to measure distortion we shall take $X = \ell_q^N$
- Recall our classes $\mathcal{A}^s(X)$, $s > 0$ consisting of signals for which $\sigma_k(x)_X \leq M k^{-s}$
- if $X = \ell_q^N$ the $\mathcal{A}^s = w\ell_p^N$, $p = (s + 1/q)^{-1}$

Model II: Compressible Signals

- The sparse signal classes do not represent real signals: signals will typically have all entries nonzero but most will be small
- A compressible signal x is one that can be approximated well by elements from Σ_k:
  $\sigma_k(x)_X := \inf_{z\in\Sigma_k} \|x - z\|_X$
- to measure distortion we shall take $X = \ell_q^N$
- Recall our classes $\mathcal{A}^s(X)$, $s > 0$ consisting of signals for which $\sigma_k(x)_X \leq M k^{-s}$
- if $X = \ell_q^N$ the $\mathcal{A}^s = w\ell_p^N$, $p = (s + 1/q)^{-1}$
- Closely related signal classes are balls in ℓ_p^N

Optimality on Models II

- With model classes II, we will not be able to exactly reproduce the signal with n samples since this is incomplete information

Optimality on Models II

- With model classes II, we will not be able to exactly reproduce the signal with n samples since this is incomplete information
- We are interested in what is the best we can do with the budget n with distortion measured in our chosen metric $\|\cdot\|_{\ell_q^N}$

Optimality on Models II

- With model classes II, we will not be able to exactly reproduce the signal with n samples since this is incomplete information
- We are interested in what is the best we can do with the budget n with distortion measured in our chosen metric $\|\cdot\|_{\ell_q^N}$
- Let K be a compact set in $\mathbb{R}^N$

Optimality on Models II

- With model classes II, we will not be able to exactly reproduce the signal with n samples since this is incomplete information
- We are interested in what is the best we can do with the budget n with distortion measured in our chosen metric $\|\cdot\|_{\ell_q^N}$
- Let K be a compact set in $\mathbb{R}^N$
- Best performance on the class K in norm $\|\cdot\|_X$

$$E_n(K)_X := \inf_{(\Phi,\Delta)\in\mathcal{A}_n} \sup_{x\in K} \|x - \Delta(\Phi(x))\|_X$$

Optimality on Models II

- With model classes II, we will not be able to exactly reproduce the signal with n samples since this is incomplete information
- We are interested in what is the best we can do with the budget n with distortion measured in our chosen metric $\|\cdot\|_{\ell_q^N}$
- Let K be a compact set in $\mathbb{R}^N$
- Best performance on the class K in norm $\|\cdot\|_X$

$$E_n(K)_X := \inf_{(\Phi,\Delta)\in\mathcal{A}_n} \sup_{x\in K} \|x - \Delta(\Phi(x))\|_X$$

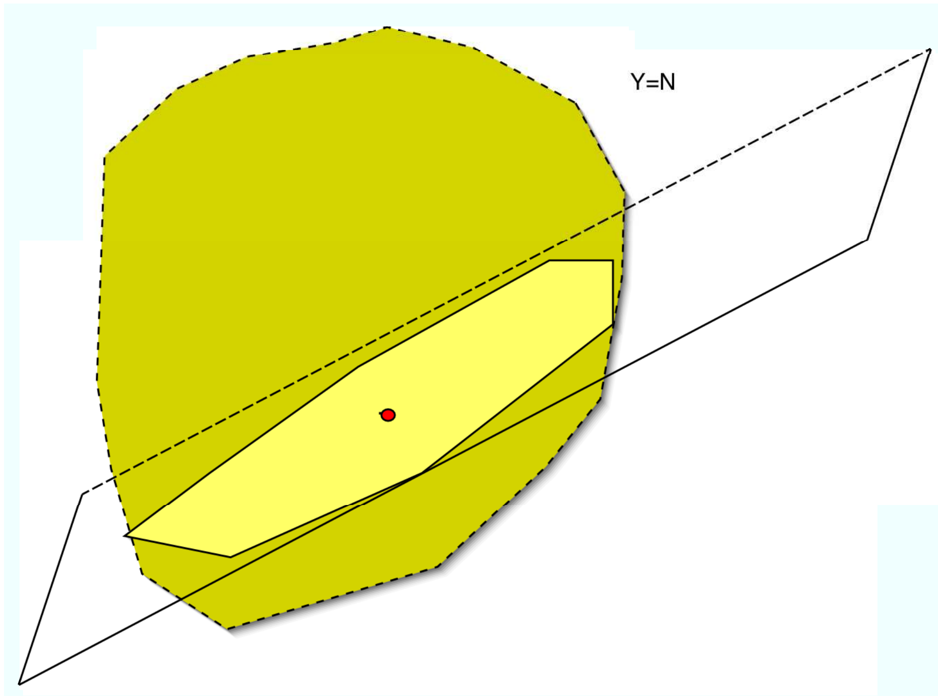
- Here $\mathcal{A}_n = \{(\Phi, \Delta) : \ \Phi \text{ is n} \times \text{N}\}$

Gelfand widths

- $E_{n,N}(K)_X$ is closely related to Gelfand widths

$$d^n(K)_X := \inf_{\mathrm{codim}(Y)=n} \sup_{x \in K \cap Y} \|x\|_X$$

The set $K \cap Y$



Gelfand widths

- $E_{n,N}(K)_X$ is closely related to Gelfand widths
  $d^n(K)_X := \inf_{\text{codim}(Y)=n} \sup_{x \in K \cap Y} \|x\|_X$
- Theorem: If $K = -K$ and $K + K \subset CK$, then

$$d^n(K)_X \leq E_{n,N}(K)_X \leq C d^n(K)_X$$

Gelfand widths

- $E_{n,N}(K)_X$ is closely related to Gelfand widths $d^n(K)_X := \inf_{\mathrm{codim}(Y)=n} \sup_{x \in K \cap Y} \|x\|_X$
- Theorem: If $K = -K$ and $K + K \subset CK$, then

$$d^n(K)_X \leq E_{n,N}(K)_X \leq C d^n(K)_X$$

- $Y \leftrightarrow \mathcal{N}$

Proof of Theorem

- Lower inequality: Suppose (Φ, Δ) gives bound $E_n(K)$. Choose $Y := \mathcal{N}$. If $x \in K \cap Y$ then

$$\|x\|_X \leq \max(\|x - \Delta(0)\|_X, \| - x - \Delta(0)\|_X) \leq E_n(K)_X$$

Proof of Theorem

- Lower inequality: Suppose (Φ, Δ) gives bound $E_n(K)$. Choose $Y := \mathcal{N}$. If $x \in K \cap Y$ then

$$\|x\|_X \leq \max(\|x - \Delta(0)\|_X, \| - x - \Delta(0)\|_X) \leq E_n(K)_X$$

- Upper inequality: Let Y be an optimal space for Gelfand width. Take an orthonormal basis for $Y^\perp$ and use these as the rows of Φ. Choose a decoder Δ so that $\Delta(y)$ is in K whenever $K \cap \mathcal{F}(y) \neq 0$. Now if $x \in K$ then $x - \Delta(\Phi(x))$ is in $K - K \subset CK$. It is also in $\mathcal{N}$. Hence $C^{-1}(x - \Delta(\Phi(x))$ is in $K \cap \mathcal{N}$. Hence

$$C^{-1}\|x - \Delta(\Phi(x))\| \leq d^n(K)_C$$

Optimality on ℓ_p^N classes

- The asymptotic behavior of $E_n(K)_X$ is known for all $K = U(\ell_p^N)$ in all $X = \ell_q^N$

Optimality on ℓ_p^N classes

- The asymptotic behavior of $E_n(K)_X$ is known for all $K = U(\ell_p^N)$ in all $X = \ell_q^N$
- Solved in 1970's and 1980's in Approximation Theory and Finite Dimensional Geometry

Optimality on ℓ_p^N classes

- The asymptotic behavior of $E_n(K)_X$ is known for all $K = U(\ell_p^N)$ in all $X = \ell_q^N$
- Solved in 1970's and 1980's in Approximation Theory and Finite Dimensional Geometry
- Kashin, Gluskin main players

Optimality on ℓ_p^N classes

- The asymptotic behavior of $E_n(K)_X$ is known for all $K = U(\ell_p^N)$ in all $X = \ell_q^N$
- Solved in 1970's and 1980's in Approximation Theory and Finite Dimensional Geometry
- Kashin, Gluskin main players
- For $K = U(\ell_p^N)$, we have

$$C_1\Psi(n, N, p, q) \le d^n(K)_{\ell_p} \le C_2\Psi(n, N, p.q),$$

Optimality on ℓ_p^N classes

- The asymptotic behavior of $E_n(K)_X$ is known for all $K = U(\ell_p^N)$ in all $X = \ell_q^N$
- Solved in 1970's and 1980's in Approximation Theory and Finite Dimensional Geometry
- Kashin, Gluskin main players
- For $K = U(\ell_p^N)$, we have

$$C_1\Psi(n, N, p, q) \le d^n(K)_{\ell_p} \le C_2\Psi(n, N, p.q),$$

- C_1, C_2 only depend on p and q, and

Optimality on ℓ_p^N classes

- The asymptotic behavior of $E_n(K)_X$ is known for all $K = U(\ell_p^N)$ in all $X = \ell_q^N$
- Solved in 1970's and 1980's in Approximation Theory and Finite Dimensional Geometry
- Kashin, Gluskin main players
- For $K = U(\ell_p^N)$, we have

$$C_1 \Psi(n, N, p, q) \leq d^n(K)_{\ell_p} \leq C_2 \Psi(n, N, p.q),$$

- C_1, C_2 only depend on p and q, and
- $\Psi(n, N, p, q) := [\min(1, N^{1-1/p} n^{-1/2})]^{\frac{1/p-1/q}{1/p-1/2}}, \quad 1 \leq n \leq N, \ 1 < p < q \leq 2$

Remarks

- The hardest case is $p = 1, q = 2$

Remarks

- The hardest case is $p = 1, q = 2$
- $C_0\sqrt{\frac{\log(N/n)}{n}} \leq E_n(U(\ell_1^N))_{\ell_2^N} \leq C_1\sqrt{\frac{\log(N/n)}{n}}$

Remarks

- The hardest case is $p = 1, q = 2$
- $C_0\sqrt{\frac{\log(N/n)}{n}} \leq E_n(U(\ell_1^N))_{\ell_2^N} \leq C_1\sqrt{\frac{\log(N/n)}{n}}$
- Kashin proved the upper bound with a slightly inferior logarithm

Remarks

- The hardest case is $p = 1, q = 2$
- $C_0\sqrt{\frac{\log(N/n)}{n}} \leq E_n(U(\ell_1^N))_{\ell_2^N} \leq C_1\sqrt{\frac{\log(N/n)}{n}}$
- Kashin proved the upper bound with a slightly inferior logarithm
- The matrices he used were to fill in the entries with independent draws of a Bernouli random variable

Remarks

- The hardest case is $p = 1, q = 2$
- $C_0\sqrt{\frac{\log(N/n)}{n}} \leq E_n(U(\ell_1^N))_{\ell_2^N} \leq C_1\sqrt{\frac{\log(N/n)}{n}}$
- Kashin proved the upper bound with a slightly inferior logarithm
- The matrices he used were to fill in the entries with independent draws of a Bernouli random variable
- Gluskin improved the upper bound and proved the lower bound (with Garneev). He used matrices whose entries are independent draws of a Gaussian matrix

Remarks

- The hardest case is $p = 1, q = 2$
- $C_0\sqrt{\frac{\log(N/n)}{n}} \leq E_n(U(\ell_1^N))_{\ell_2^N} \leq C_1\sqrt{\frac{\log(N/n)}{n}}$
- Kashin proved the upper bound with a slightly inferior logarithm
- The matrices he used were to fill in the entries with independent draws of a Bernouli random variable
- Gluskin improved the upper bound and proved the lower bound (with Garneev). He used matrices whose entries are independent draws of a Gaussian matrix
- It was though for a long time that Bernouli did not work

Remarks

- The hardest case is $p = 1, q = 2$
- $C_0\sqrt{\frac{\log(N/n)}{n}} \leq E_n(U(\ell_1^N))_{\ell_2^N} \leq C_1\sqrt{\frac{\log(N/n)}{n}}$
- Kashin proved the upper bound with a slightly inferior logarithm
- The matrices he used were to fill in the entries with independent draws of a Bernouli random variable
- Gluskin improved the upper bound and proved the lower bound (with Garneev). He used matrices whose entries are independent draws of a Gaussian matrix
- It was though for a long time that Bernouli did not work
- These theorems were thought to be very deep at the time. We shall give rather simple proofs of them using compressed sensing.

Interpretation of these results

- Let us begin by comparing the above results with the more classical methods of compression

Interpretation of these results

- Let us begin by comparing the above results with the more classical methods of compression
- Keep $p = 1$, $q = 2$ for simplicity

Interpretation of these results

- Let us begin by comparing the above results with the more classical methods of compression
- Keep $p = 1$, $q = 2$ for simplicity
- Given a value of k if we take n so that $n = k \log(N/n)$ then we obtain for any $x \in \ell_1$ (also holds for $x \in w\ell_1$)

$$\|x - \Delta(\Phi(x)\|_{\ell_2^N} \leq C\sqrt{\frac{\log(N/n)}{n}} = C\frac{1}{\sqrt{k}}$$

Interpretation of these results

- Let us begin by comparing the above results with the more classical methods of compression
- Keep $p = 1$, $q = 2$ for simplicity
- Given a value of k if we take n so that $n = k\log(N/n)$ then we obtain for any $x \in \ell_1$ (also holds for $x \in w\ell_1$)

$$\|x - \Delta(\Phi(x)\|_{\ell_2^N} \leq C\sqrt{\frac{\log(N/n)}{n}} = C\frac{1}{\sqrt{k}}$$

- In other words we get the same performance as if we took the N measurements (of the coordinates of x and then retained the largest k of them

Interpretation of these results

- Let us begin by comparing the above results with the more classical methods of compression
- Keep $p = 1$, $q = 2$ for simplicity
- Given a value of k if we take n so that $n = k \log(N/n)$ then we obtain for any $x \in \ell_1$ (also holds for $x \in w\ell_1$)

$$\|x - \Delta(\Phi(x)\|_{\ell_2^N} \leq C\sqrt{\frac{\log(N/n)}{n}} = C\frac{1}{\sqrt{k}}$$

- In other words we get the same performance as if we took the N measurements (of the coordinates of x and then retained the largest k of them
- This is the advantage of compressed sensing: the number of measurements we need to make is only slightly larger than k

Better Performance

- The above remarks motivate the following

Better Performance

- The above remarks motivate the following
- We say (Φ, Δ) is Instance-Optimal of order k for X if for an absolute constant $C > 0$ (independent of k, n, N)

$$\|x - \Delta(\Phi(x))\|_X \leq C\sigma_k(x)_X$$

Better Performance

- The above remarks motivate the following
- We say (Φ, Δ) is Instance-Optimal of order k for X if for an absolute constant $C > 0$ (independent of k, n, N)

$$\|x - \Delta(\Phi(x))\|_X \leq C\sigma_k(x)_X$$

- We will be interested in $X = \ell_q^N$

Better Performance

- The above remarks motivate the following
- We say (Φ, Δ) is Instance-Optimal of order k for X if for an absolute constant $C > 0$ (independent of k, n, N)

$$\|x - \Delta(\Phi(x))\|_X \le C\sigma_k(x)_X$$

- We will be interested in $X = \ell_q^N$
- Problem: for a given X and size $n \times N$ find the largest values of k for which we have instance-optimality and the encoder-decoder pairs (Φ, Δ) which admit these values of k

Optimal Matrices

- What properties of a matrix Φ determine whether it is optimal or near optimal for these problems?

Optimal Matrices

- What properties of a matrix Φ determine whether it is optimal or near optimal for these problems?
- The secret lies in the null space $\mathcal{N}$ of Φ

Optimal Matrices

- What properties of a matrix Φ determine whether it is optimal or near optimal for these problems?
- The secret lies in the null space $\mathcal{N}$ of Φ
- We say Φ has the null space property in X if

$$\|\eta\|_X \leq C_0 \sigma_k(\eta)_X, \quad \eta \in \mathcal{N}$$

Optimal Matrices

- What properties of a matrix Φ determine whether it is optimal or near optimal for these problems?
- The secret lies in the null space $\mathcal{N}$ of Φ
- We say Φ has the null space property in X if

$$\|\eta\|_X \leq C_0 \sigma_k(\eta)_X, \quad \eta \in \mathcal{N}$$

- When $X = \ell_q^N$ for some q then an equivalent formulation is

$$\|\eta_T\|_X \leq C_1 \|\eta_{T^c}\|_X, \quad \eta \in \mathcal{N}$$

Optimal Matrices

- What properties of a matrix Φ determine whether it is optimal or near optimal for these problems?
- The secret lies in the null space $\mathcal{N}$ of Φ
- We say Φ has the null space property in X if

$$\|\eta\|_X \leq C_0 \sigma_k(\eta)_X, \quad \eta \in \mathcal{N}$$

- When $X = \ell_q^N$ for some q then an equivalent formulation is

$$\|\eta_T\|_X \leq C_1 \|\eta_{T^c}\|_X, \quad \eta \in \mathcal{N}$$

- Elements in the null space should have no structure - look like noise

Main Result

Theorem (Cohen-Dahmen-DeVore) Given an $n \times N$ matrix Φ, a norm $\| \cdot \|_X$ and a value of k, then to have instance optimality in X with a constant C_0 a necessary and sufficient is that Φ has the null space of order $2k$ with a constant C_1 where $C_1 = C_0/2$ in the sufficient part and $C_1 = C_0$ in the necessary part.

Proof of Sufficiency

- define a decoder Δ for Φ

$$\Delta(y) := \operatorname*{Argmin}_{z \in \mathcal{F}(y)} \sigma_k(z)_X$$

Proof of Sufficiency

- define a decoder Δ for Φ

$$\Delta(y) := \operatorname*{Argmin}_{z \in \mathcal{F}(y)} \sigma_k(z)_X$$

- $\eta := x - \Delta(\Phi(x))$ is in $\mathcal{N}$

Proof of Sufficiency

- define a decoder Δ for Φ

$$\Delta(y) := \operatorname*{Argmin}_{z \in \mathcal{F}(y)} \sigma_k(z)_X$$

- $\eta := x - \Delta(\Phi(x))$ is in $\mathcal{N}$
- $\|x - \Delta(\Phi(x))\|_X \leq (C_0/2)\sigma_{2k}(x - \Delta(\Phi(x))_X$
  $\leq (C_0/2)(\sigma_k(x)_X + \sigma_k(\Delta(\Phi(x))_X) \leq C_0\sigma_k(x)_X$

Proof of Sufficiency

- define a decoder Δ for Φ

$$\Delta(y) := \operatorname*{Argmin}_{z\in\mathcal{F}(y)} \sigma_k(z)_X$$

- $\eta := x - \Delta(\Phi(x))$ is in $\mathcal{N}$
- $\|x - \Delta(\Phi(x))\|_X \le (C_0/2)\sigma_{2k}(x - \Delta(\Phi(x))_X$
  $\le (C_0/2)(\sigma_k(x)_X + \sigma_k(\Delta(\Phi(x))_X) \le C_0\sigma_k(x)_X$
- We used $\sigma_{2k}(x+z)_X \le \sigma_k(x)_X + \sigma_k(z)_X$

Proof of Sufficiency

- define a decoder Δ for Φ

$$\Delta(y) := \operatorname*{Argmin}_{z \in \mathcal{F}(y)} \sigma_k(z)_X$$

- $\eta := x - \Delta(\Phi(x))$ is in $\mathcal{N}$
- $\|x - \Delta(\Phi(x))\|_X \leq (C_0/2)\sigma_{2k}(x - \Delta(\Phi(x))_X$
  $\leq (C_0/2)(\sigma_k(x)_X + \sigma_k(\Delta(\Phi(x))_X) \leq C_0\sigma_k(x)_X$
- We used $\sigma_{2k}(x+z)_X \leq \sigma_k(x)_X + \sigma_k(z)_X$
- The last inequality uses the fact that $\Delta(\Phi(x))$ minimizes $\sigma_k(z)$ over $\mathcal{F}(y)$