Homework 8

Assigned: 14/11/2011.
Due: 25/11/2011.

Exercise 1. Asymptotics of best relative k-term approximation error
Undetermined linear regression (ULR) is a fundamental problem with broad applications in many fields. In ULR we seek an unknown vector $\mathbf{x} \in \mathbb{R}^N$ from its dimensionality reducing linear projection $\mathbf{y} \in \mathbb{R}^m$ $(m < N)$ obtained via a known encoding matrix $\Phi \in \mathbb{R}^{m \times N}$ as:

$$\mathbf{y} = \Phi \mathbf{x}.$$

Since Φ posesses a non-trivial kernel, we clearly need to make additional assumptions about $\mathbf{x}$ to distinguish it from the infinitely many possible solutions. It is now well known that a *sparsity* assumption on $\mathbf{x}$, i.e. $\mathbf{x}$ has most of its energy in $k \ll N$ coefficients, plays a crucial role in obtaining "good" solutions. Furthermore from a probabilistic perspective one assumes $\mathbf{x}$ to be drawn from a *prior*. Hence in this probabilistic setting, own would like to come up with a mathematically precise mechanism to characterize the "compressibility" of a prior.

In this exercise we will analyze the relative best k - term approximation error:

$$\bar{\sigma}_k(\mathbf{x})_q = \frac{\sigma_k(\mathbf{x})_q}{\|\mathbf{x}\|_q}$$

in the limit of large problem sizes (i.e. $N \to \infty$), for any $q \in (0, \infty)$ and for random vectors $\mathbf{x}_N = (X_1, \ldots, X_n)$ with i.i.d entries ($X_i \sim p(x)$) drawn from a distribution $p(x)$. Note that $\sigma_k(\mathbf{x})_q = \inf_{\|z\|_0 \le k} \|\mathbf{x} - \mathbf{z}\|_q$. Denote $\tilde{p}(x)$ as the PDF of $\tilde{X}_n = |X|_n$ and $\tilde{F}(t) := \mathbb{P}(|X| \le t)$ as its CDF. Assume $\tilde{F}$ is continuous and strictly increasing. Also assume that X_n satisfies $\mathbb{E}\,|X|^q < \infty$. For any $\kappa \in (0,1)$ consider the following function:

$$G_q[p](\kappa) := \frac{\int_0^{\tilde{F}^{-1}(1-\kappa)} x^q \tilde{p}(x) dx}{\int_0^{\infty} x^q \tilde{p}(x) dx}.$$

Given any sequence $(k_N)_{N=1}^{\infty}$ so that $\lim_{N\to\infty} \frac{k_N}{N} = \kappa$, we would like to show in this question that:

$$\lim_{N\to\infty} \bar{\sigma}_{k_N}(\mathbf{x}_N)_q^q = G_q[p](\kappa) \quad \text{a.s} \tag{1}$$

where a.s. denotes "almost surely".

Let us begin by defining the random variables $Y_n := |X_n|^q$. They have the CDF $F_Y(y) := \mathbb{P}(Y \le y) = \tilde{F}(y^{1/q})$. Denote

$$\mu = \mathbb{E}[Y] = \int_0^{\infty} x^q d\tilde{F}(x)$$

1. **[1 point]** For a given $\kappa \in (0,1)$, show that there is a unique $\tau_0 \in (0,\infty)$ such that $\kappa = 1 - F_Y(\tau_0)$.

2. Fix $0 < \epsilon < \tau_0$. Define $\tau = \tau_0 - \epsilon$ and $\rho = \int_0^\tau y d\tilde{F}(y)$. Clearly $\rho \in (0, \mu)$. Lastly consider $L_N = \max\{l \leq N : \sum_{i=1}^{l} Y_{i,N} \leq N\rho\}$ where $Y_{1,N} \leq \cdots \leq Y_{N,N}$ are the increasing order statistics of $Y_1, \ldots, Y_N$. It can be shown that

$$\lim_{N\to\infty} \frac{L_N}{N} = F_Y(\tau) \quad \text{a.s.} \tag{2}$$

We now proceed to prove (1) by solving the following questions:

(a) **[1 point]** Show that: $\lim_{N\to\infty} \frac{N-k_N}{L_N} > 1$.

(b) **[3 points]** Using the above result and applying (2) show that:

$$\lim \inf_{N\to\infty} \frac{\sigma_{k_N}(\mathbf{x}_N)_q^q}{\|\mathbf{x}\|_q^q} \geq \frac{\int_0^{\tau_0-\epsilon} y dF_Y(y)}{\int_0^\infty y dF_Y(y)} \quad \text{a.s.}$$

3. **[2 points]** By now choosing $\tau = \tau_0 + \epsilon$ (ϵ is the same as in the previous part), follow the steps of the previous part to show that:

$$\lim \sup_{N\to\infty} \frac{\sigma_{k_N}(\mathbf{x}_N)_q^q}{\|\mathbf{x}\|_q^q} \leq \frac{\int_0^{\tau_0+\epsilon} y dF_Y(y)}{\int_0^\infty y dF_Y(y)} \quad \text{a.s.}$$

4. **[3 points]** Now finally by using the results of the last two parts, deduce that:

$$\lim_{N\to\infty} \frac{\sigma_{k_N}(\mathbf{x}_N)_q^q}{\|\mathbf{x}\|_q^q} = G_q[p](\kappa) \quad \text{a.s}$$

Exercise 2. Least squares estimation versus Oracle k-sparse Estimation
Carrying on from the previous problem, a natural question one could ask is the following. Given that the data $\mathbf{x}_N = (X_1, \ldots, X_N)$ is formed from i.i.d samples from some distribution $p(x)$ ($X_i \sim p(x)$), then how would one characterize the compressibility of the data $\mathbf{x}_N$? A possible way to do this would be to apply a dimensionality reducing linear operator on $\mathbf{x}_N$ and then compare the relative k term approximation error performance of a "sparse estimator" with a typically "dense" (or non-sparse estimator).

In this exercise we will compare the expected performance of two decoding approaches for estimating a given vector $\mathbf{x} \in \mathbb{R}^N$ from its encoding : $\mathbf{y} = \Phi\mathbf{x}$. Here, Φ is a $m \times N$ matrix (m < N) with i.i.d Gaussian entries, $\phi_{i,j} \sim \mathcal{N}(0, \frac{1}{m})$. In order to estimate $\mathbf{x}$ from the measurement $\mathbf{y}$, we will compare two decoding approaches:

1. **Oracle k sparse decoder**

$$\triangle_{oracle}(\mathbf{y}, \Lambda) = \arg\min_{\tilde{\mathbf{x}}:\text{support}(\tilde{\mathbf{x}})=\Lambda} \|\mathbf{y} - \Phi\tilde{\mathbf{x}}\|_2$$

$\triangle_{oracle}$ is an idealized sparse decoder which is given the "side" information Λ corresponding to the indices of the k largest coefficients of $\mathbf{x}$ (Assume $k < m$). Note that the estimation $\triangle_{oracle}(\mathbf{y}, \Lambda)$ has at most k non-zero coefficients.

2. **Least squares (LS) decoder**

$$\triangle_{LS}(\mathbf{y}) = \arg\min_{\tilde{\mathbf{x}}:\mathbf{y}=\Phi\tilde{\mathbf{x}}} \|\tilde{\mathbf{x}}\|_2$$

This is a commonly used decoder which typically provides a "dense" estimate of $\mathbf{x}$ due to the particular nature of the objective function (l_2 norm).

In the following, for any column index set $\Lambda \subset \{1, \ldots, N\}$ we have the notation that Φ_Λ is the matrix restricted to the column set Λ. Similarly for any $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{x}_\Lambda$ denotes the restriction of $\mathbf{x}$ to the support set Λ. The complement of Λ is denoted by $\bar{\Lambda}$.

1. **[3 points]** Show that :

$$\triangle_{LS}(\mathbf{y}) = \Phi^{+}(\mathbf{y})$$

where $\Phi^+ = \Phi^T(\Phi\Phi^T)^{-1}$ is the pseudo inverse of Φ. Also show that:

$$\triangle_{oracle}(\mathbf{y}, \Lambda) = \Phi_\Lambda^{+}(\mathbf{y})$$

where Φ_Λ is a $m \times k$ matrix and $\Phi_\Lambda^+ = (\Phi_\Lambda^T \Phi_\Lambda)^{-1}\Phi_\Lambda^T$ is the pseudo inverse of Φ_Λ.

2. It is well known that the relative expected performance of $\triangle_{LS}$ is given by:

$$\frac{\mathbb{E}_\Phi[\|\triangle_{LS}(\Phi\mathbf{x}) - \mathbf{x}\|_2^2]}{\|\mathbf{x}\|_2^2} = 1 - \frac{m}{N}.$$

Observe that the expected performance of $\triangle_{LS}$ is directly governed by the *undersampling ratio*: $\frac{m}{N}$. It is independent of the vector $\mathbf{x}$ which should be no surprise since the Gaussian distribution is isotropic. We now proceed to derive the expression for the relative expected performance of $\triangle_{oracle}$. We will see that the expected performance of the oracle estimator drastically depends on the shape of the best k term approximation relative error of $\mathbf{x}$ (i.e. $\bar{\sigma}_k(\mathbf{x})$).

   (a) **[2 points]** Denoting $w := \dfrac{\Phi_{\bar{\Lambda}}\mathbf{x}_{\bar{\Lambda}}}{\|\Phi_{\bar{\Lambda}}\mathbf{x}_{\bar{\Lambda}}\|_2} \in \mathbb{R}^m$ show that:

$$\frac{\|\triangle_{oracle}(\mathbf{y}, \Lambda) - \mathbf{x}\|_2^2}{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2} = \left\|\Phi_\Lambda^+ w\right\|_2^2 \cdot \frac{\|\Phi_{\bar{\Lambda}}\mathbf{x}_{\bar{\Lambda}}\|_2^2}{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2} + 1.$$

   (b) **[2 points]** Show that: $\mathbb{E}_\Phi\left[\dfrac{\|\Phi_{\bar{\Lambda}}\mathbf{x}_{\bar{\Lambda}}\|_2^2}{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2}\right] = 1.$

   (c) **[2 points]** Let $\Phi_\Lambda = U\Sigma V^T$ be the SVD of Φ_Λ where u_l and v_l denote the column vectors of U and V repectively. It can be shown that:

   - The random variables $\langle u_l, w\rangle$ are identically distributed and statistically independent from Φ_Λ.
   - $\mathbb{E}[|\langle u_l, w\rangle|^2] = \dfrac{1}{m}$. Furthermore, $\mathbb{E}[\text{Trace}(\Phi_\Lambda^T\Phi_\Lambda)^{-1}] = \dfrac{mk}{m-k+1}$.

   With help from the above facts show that: $\mathbb{E}[\left\|\Phi_\Lambda^+ w\right\|_2^2] = \dfrac{k}{m-k+1}$.

   (d) **[2 points]** Lastly using the above results conclude that:

$$\frac{\mathbb{E}[\|\triangle_{oracle}(\mathbf{y}, \Lambda) - \mathbf{x}\|_2^2]}{\|\mathbf{x}\|_2^2} = \frac{1}{1 - \dfrac{k}{m+1}} \frac{\sigma_k(\mathbf{x})_2^2}{\|\mathbf{x}\|_2^2}$$