Infotheory for Statistics and Learning

Lecture 11

- Sparse denoising [PW:30.2]
- Sparse linear regression [PW:30.2],[RWY]
- Compressed sensing [CRT]
- Almost lossless analog compression [WV]

Notation for asymptotic behavior:

$f(n) = \Theta(g(n)) \iff$ there is an $n_0 > 0$ and constants C_1, C_2 such that for all $n > n_0$, $C_1 g(n) \le f(n) \le C_2 g(n)$

$f(n) \lesssim g(n) \iff$ there is an $n_0 > 0$ and a constant C_1 such that for all $n > n_0$, $f(n) \le C_1 g(n)$

$f(n) \gtrsim g(n) \iff$ there is an $n_0 > 0$ and a constant C_2 such that for all $n > n_0$, $f(n) \ge C_2 g(n)$

That is, $f(n) = \Theta(g(n)) \iff g(n) \lesssim f(n) \lesssim g(n)$

Sparse Denoising

Consider the GLM, $Y_i = \theta + Z_i$, where $Z_i \sim \mathcal{N}(0, I_p)$, $i = 1, \dots, n$ i.i.d. and $\theta \in \mathbb{R}^p$

Assume θ is sparse in the sense $\|\theta\|_0 \le k < p$, $\|\theta\|_0 = |\{i : \theta_i \ne 0\}|$

Let $T_k = \{\theta : \|\theta\|_0 \le k\}$ and consider the minimax risk for $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$ and $n = 1$

$$R_1^*(T_k) = \inf_{\hat{\theta}} \sup_{\theta \in T_k} E_\theta[\|\theta - \hat{\theta}(Y)\|^2]$$

where E_θ denotes expectation over $Y = \theta + Z \sim \mathcal{N}(\theta, I_p)$

For $n > 1$ we get

$$R_n^*(T_k) = \frac{1}{n} R_1^*(T_k)$$

because $\bar{Y} = n^{-1} \sum_i Y_i$ is a sufficient statistic

Lower bound on $R^*(T_k) = R_1^*(T_k)$:

Since $R_\pi^* \le R^*$ (Bayesian $\le$ minimax) for any prior π on θ, we can choose π by drawing b uniformly from $\{b \in \{0,1\}^p : \|b\|_0 = k\}$ and setting $\theta = \tau b$ for some $\tau > 0 \Rightarrow b \to \theta \to Y \to \hat{\theta} \to \hat{b}$

We have

$$I(\theta; \hat{\theta}) \le \sup_\pi I(\theta; Y) \le \sup_{\theta \ne \theta'} D(\mathcal{N}(\theta, I_P) \| \mathcal{N}(\theta', I_P)) \le k\tau^2$$

Assume we use $\hat{b} = \min \|\hat{\theta} - \tau b\|^2$ over $\{b \in \{0,1\}^p : \|b\|_0 = k\}$, then $\tau^2 \|b - \hat{b}\|_0 \le 4\|\theta - \hat{\theta}\|^2 \Rightarrow \tau^2 E[\|b - \hat{b}\|_0] \le 4R^*$

Thus $I(b; \hat{b}) \ge \min I(b; \hat{b})$ where the min is over distributions on b such that $E[\|b - \hat{b}\|_0] \le 4R^*/\tau^2$, leading to the bound (in nats)

$$I(b; \hat{b}) \ge \ln \binom{p}{k} - p\, h\left(\frac{4R^*}{\tau^2 p}\right)$$

where $h(x) = -x \ln x - (1-x)\ln(1-x)$

Since $E[\|\theta - \hat{\theta}\|^2] \le E[\|\theta\|^2] = k\tau^2$ we can set $R^*_\pi = \varepsilon(k) k\tau^2$ with $\varepsilon(k) \in (0,1)$, and since $I(b;\hat{b}) \le I(\theta;\hat{\theta})$ we get

$$\ln \binom{p}{k} - p\, h\left(\frac{4\varepsilon(k)k}{p}\right) \le \frac{R^*_\pi}{\varepsilon(k)} \le \frac{R^*}{\varepsilon(k)}$$

Now assume $k = k(p) \to \infty$ as $p \to \infty$. Then Stirling $\Rightarrow$

$$\ln \binom{p}{k} \approx \frac{1}{2} \ln \frac{p}{k(p-k)2\pi} + p\, h\left(\frac{k}{p}\right)$$

Assuming $\varepsilon_0 < \varepsilon(k) < (1-\varepsilon_0)/4$, for some $0 < \varepsilon_0 \ll 1$, and $k/p < 1/2 \Rightarrow$

$$R^* > \varepsilon_0 \left[\frac{1}{2} \ln \frac{p}{k(p-k)2\pi} + p\, h\left(\frac{k}{p}\right) - p\, h\left(\frac{(1-\varepsilon_0)k}{p}\right)\right]$$
$$\gtrsim p\, h\left(\frac{k}{p}\right) \gtrsim k \ln \frac{p}{k}$$

Upper bound on $R^*(T_k)$

For $Y = \theta + Z$ we study $\hat{\theta} = \arg\min_{\theta \in T_k} \|Y - \theta\|^2$

We get (with $\cdot$ = scalar product)
$\|Z - (\hat{\theta} - \theta)\|^2 \le \|Y - \theta\|^2 = \|Z\|^2 \Rightarrow \|\theta - \hat{\theta}\|^2 \le 2(\theta - \hat{\theta}) \cdot Z$

Consequently, since also $\|\theta - \hat{\theta}\|_0 \le 2k$,

$$\frac{1}{2}\|\theta - \hat{\theta}\| \le \sup_{u \in S^p \cap T_{2k}} Z \cdot u = \max_{|J| = 2k} \|Z_J\|$$

where S^p = the unit sphere in $\mathbb{R}^p$, Z_J the sub-vector defined by J

Because $Z \sim \mathcal{N}(0, I_p)$, it can now be shown that

$$\Pr\left(\|Z_J\|^2 \ge k \ln \frac{pe}{k}\right) \le \exp\left(-\frac{ck}{2} \ln \frac{p}{k}\right)$$

$\Rightarrow$ for $\ell = k\ln(p/k)$ and $\varepsilon > 0$, there is an L such that for $\ell > L$

$$\Pr\left(\|Z_J\|^2 \geq k\ln\frac{pe}{k}\right) \leq \varepsilon$$

Hence for large ℓ

$$E[\|\theta - \hat{\theta}\|^2] \leq 4k\ln\frac{pe}{k} = \Theta\left(k\ln\frac{p}{k}\right)$$

That is,

$$R^* \lesssim k\ln\frac{p}{k}$$

Consequently

$$R^* = \Theta\left(k\ln\frac{p}{k}\right)$$

Sparse Linear Regression

$Y = X\theta + Z$, $Y \in \mathbb{R}^{n\times 1}$, $\theta \in T_k \subset \mathbb{R}^{p\times 1}$, $n \geq p$, $k < p$,

$X_{ij} \sim \mathcal{N}(0, 1/n)$ and independent; $Z \sim \mathcal{N}(0, I_n)$

For $\hat{\theta} = \hat{\theta}(X, Y)$ the minimax risk is

$$R^* = R_n^*(T_k) = \inf_{\hat{\theta}} \sup_{\theta\in T_k} E_\theta\|\theta - \hat{\theta}(X,Y)\|^2$$

with E_θ over X and $Y \sim \mathcal{N}(0, (\|\theta\|^2/n + 1)I_n)$

Bounding $I(\theta; \hat{\theta})$ it can be shown that

$$R^* \gtrsim k\ln\frac{p}{k}$$

for any n; i.e. the same lower bound as for $n = p$ and $X = I_p$

To get an upper bound, consider $\hat{\theta} = \arg\min_{\theta \in T_k} \|Y - X\theta\|^2$

$\Rightarrow \|X(\theta - \hat{\theta})\|^2 \leq 2\|\theta - \hat{\theta}\| \sup_{u \in S^p \cap T_{2k}} Z \cdot (Xu)$

For $J = \{i : (\theta - \hat{\theta})_i \neq 0\}$, let X_J be the corresponding part of X

Then with $v = \theta - \hat{\theta}$

$$\inf_{v \in T_{2k}} \frac{\|Xv\|}{\|v\|} = \min_{|J| \leq 2k} \sigma_{\min}(X_J)$$

where $\sigma_{\min}(X_J)$ is the smallest singular value of X_J

For $\ell = k \ln(p/k)$, $\Pr(\min_{|J| \leq 2k} \sigma_{\min}(X_J) < 1/2) \to 0$ as $\ell \to \infty$

$\Rightarrow \|\theta - \hat{\theta}\| < 2\|X(\theta - \hat{\theta})\|$ with high prob. as $\ell \to \infty$

Now, similarly as for $n = p$ and $X = I_p$, we can show that

$$\sup_{u \in S^p \cap T_{2k}} Z \cdot (Xu) \lesssim \sqrt{k \ln \frac{ep}{k}}$$

with high probability, so overall we have

$$R^* \lesssim k \ln \frac{p}{k}$$

Compressed Sensing

Consider $y = X\theta + z$, $y \in \mathbb{R}^{n \times 1}$, $\theta \in T_k \subset \mathbb{R}^{p \times 1}$, $k < p$ and $n < p$ (or $n \ll p$); the system is seemingly underdetermined, but $\theta \in T_k$

The elements of y are linearly compressed measurements of θ, disturbed by z

All variables are deterministic and it is known that $\|z\| \leq \varepsilon$

For $\varepsilon = 0$ a brute force approach to recovering θ from y is to try to solve all possible systems $y = X_J \theta_J$ for all J s.t. $|J| \leq k$

$\Rightarrow$ an integer program of exponential complexity

However, it turns out that we can instead solve the convex program

$$\min \|\theta\|_1 \text{ s.t. } X\theta = y$$

where $\|\theta\|_1 = \sum |\theta_i|$. Let $\tilde{\theta}$ denote the solution

Uniform uncertainty or restricted isometry:

Define $\delta_k = \delta_k(X)$ as the smallest $\delta > 0$ such that

$$(1-\delta)\|b\|^2 \le \|X_J b\|^2 \le (1+\delta)\|b\|^2$$

for all $J \subset \{1,\dots,p\}$, $|J| \le k$, and $b \in \mathbb{R}^{|J|}$

For $\varepsilon = 0$, it has been shown[1] that $\tilde{\theta} = \theta$ as long as X fulfills

$$\delta_k + \delta_{2k} + \delta_{3k} < 1$$

In the case $\varepsilon > 0$ we can instead solve the convex program

$$\min \|\theta\|_1 \;\; \text{s.t.} \;\; \|X\theta - y\| \le \varepsilon$$

Let $\hat{\theta}$ denote the solution

[1] Candès & Tao, "Decoding by linear programming," *IEEE Trans. IT*, Dec. 2005

We have the following result (see [CRT]):

As long as $\delta_{3k} + 3\delta_{4k} < 2$, $\hat{\theta}$ fulfills

$$\|\hat{\theta} - \theta\| \le C(\delta_{4k})\varepsilon$$

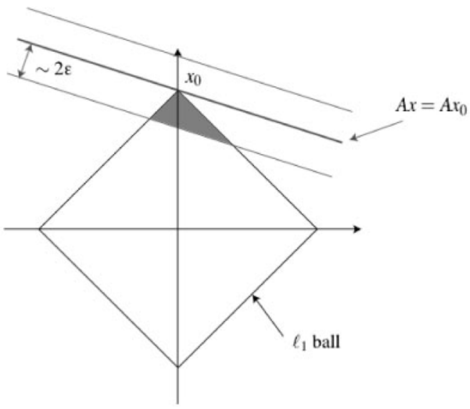

FIGURE 2.1. Geometry in $\mathbb{R}^2$. Here, the point x_0 is a vertex of the ℓ_1-ball, and the shaded area represents the set of points obeying both the tube and the cone constraints. By showing that every vector in the cone of descent at x_0 is approximately orthogonal to the null space of A, we will ensure that $x^\sharp$ is not too far from x_0.

Illustration from [CRT]

$\hat{\theta} = \theta + h \Rightarrow \|Xh\| \le 2\varepsilon$ and $\|h_{J^c}\|_1 \le \|h_J\|_1$, $J =$ support of θ, $|J| \le k$

Restricted isometry $\Rightarrow \|Xh\| \approx \|h\|$

Almost Lossless Analog Compression

In compressed sensing we had linear encoding = dimensionality reduction, $p \to n$

The general case (stochastic setting): Consider a stochastic process $\{X_i\}$ with $X_i \in \mathcal{X}$ for a given measurable space $(\mathcal{X}, \mathcal{F})$

Given another space $(\mathcal{Y}, \mathcal{G})$, an (n, k)-code for $\{X_i\}$ is, for each $1 \le k \le n < \infty$, defined by

- an encoder $f_n : \mathcal{X}^n \to \mathcal{Y}^k$
- a decoder $g_n : \mathcal{Y}^k \to \mathcal{X}^n$

where f_n is measurable in the sense $f_n^{-1}(G) \in \mathcal{F}^n$ for all $G \in \mathcal{G}^k$, and g_n is measurable in the sense $g_n^{-1}(F) \in \mathcal{G}^k$ for all $F \in \mathcal{F}^n$

Let $r(\varepsilon)$ = infimum of all r such that there is a sequence of $(n, \lfloor rn \rfloor)$-codes that fulfills

$$\limsup_{n\to\infty} \Pr(g_n(f_n(X^n)) \neq X^n) \le \varepsilon$$

Assume $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $\mathcal{F} = \mathcal{G} = \mathcal{B}$ (the Borel sets), then without further restrictions on (f_n, g_n) we have $r(\varepsilon) = 0$ for all $\varepsilon \in [0, 1]$

. . . since $(\mathbb{R}^n, \mathcal{B}^n)$ and $(\mathbb{R}, \mathcal{B})$ are Borel equivalent

However the corresponding encoders and decoders are in general highly irregular $\Rightarrow$ hard to describe and non-robust to disturbances

Assume that $\{X_i\}$ is iid with $P_X = \alpha P_c + (1-\alpha)P_d$ where P_c is abs. continuous and P_d is discrete

Then, with regularity constraints we get (see [WV]):

f_n	g_n	$r(\varepsilon)$
linear	general	α
continuous	continuous	0
general	Lipschitz	α

The decoder g_n is Lipschitz $\iff$ for every x and y in $\mathbb{R}^k$ there is an $L < \infty$ such that $\|g_n(x) - g_n(y)\| \leq L\|x - y\|$

Note that imposing that f_n and g_n be continuous does not affect $r(\varepsilon)$ (also note that continuous $\not\Leftrightarrow$ Lipschitz)

A model for sparsity

$$P_X = \alpha P_c + (1 - \alpha)\delta_0$$

where δ_0 is the Dirac measure, i.e. for $B \in \mathcal{B}$

$$\delta_x(B) = \begin{cases} 1 & x \in B \\ 0 & \text{o.w} \end{cases}$$

Then with linear encoding $r(\varepsilon) = \alpha$