On oracle inequalities related to high dimensional linear models

Yuri Golubev

CNRS, Université de Provence

Conference on Applied Inverse Problems
July 21, Vienna

Outline of the talk

1. Spectral regularization for high dimensional linear models
   - Ordered regularizations
2. The Empirical Risk Minimization
   - Excess risk penalties
3. An oracle inequality for a known noise variance
   - Short discussion
4. Unknown noise variance
   - Example: the Tikhonov-Phillips regularization

This talk deals with recovering $\theta = (\theta(1), \dots, \theta(n))^\top \in \mathbb{R}^n$ from the noisy data

$$Y = A\theta + \sigma\xi,$$

where

- A is a known $m \times n$ - matrix with $m \geq n$
- $\xi \in \mathbb{R}^n$ is a standard white Gaussian noise with

$$\mathbf{E}\xi(k)\xi(l) = \delta_{kl}, \quad k, l = 1, \dots, m$$

- *n is large (infinity)*.
- σ may be known or unknown.

Example: the linear model can be used to approximate the equation

$$y(u) = \int A(u, v)\theta(v)\, dv + \varepsilon(u).$$

Maximum likelihood estimator

The standard ML estimator is defined by

$$\hat{\theta}_0 = \arg\min_{\theta \in \mathbb{R}^n} \|Y - A\theta\|^2, \qquad \text{where} \quad \|x\|^2 = \sum_{k=1}^{m} x^2(k).$$

With a simple algebra we obtain /Moore (1920), Penrose (1955)/

$$\hat{\theta}_0 = (A^\top A)^{-1} A^\top Y.$$

Risk of the MP inversion

The risk of this inversion is computed as follows:

$$\mathbf{E}\|\hat{\theta}_0 - \theta\|^2 = \mathbf{E}\|(A^\top A)^{-1}A^\top \epsilon\|^2 = \sigma^2 \sum_{k=1}^{n} \lambda_k,$$

where λ_k are the eigenvalues of $(A^\top A)^{-1}$

$$\lambda_k A^\top A \psi_k = \psi_k, \quad \lambda_1 \leq \lambda_2, \ldots, \leq \lambda_n$$

and $\psi_k \in \mathbb{R}^n$ are the eigenvectors of $A^\top A$.

If A has a large condition number or n is large, the risk of $\hat{\theta}_0$ may be very large.

Spectral regularization

The basic idea in the spectral regularization is to suppress large λ_k in the risk of $\hat{\theta}_0$. We smooth $\hat{\theta}_0$ with the help of a properly chosen matrixes $H_\alpha, \alpha \in \mathbb{R}^+$

$$\hat{\theta}_\alpha = H_\alpha \hat{\theta}_0 = H_\alpha \left[(A^\top A)^{-1}\right](A^\top A)^{-1} A^\top Y,$$

where $H_\alpha \left[(A^\top A)^{-1}\right](s, l) = \sum_{k=1}^{n} H_\alpha(\lambda_k)\psi_s(k)\psi_l(k)$.

Typically $\lim_{\alpha \to 0} H_\alpha(\lambda) = 1, \quad \lim_{\lambda \to \infty} H_\alpha(\lambda) = 0$ for all $\alpha > 0$.

α is called regularization parameter.

Bias-variance decomposition

For the risk of $\hat{\theta}_\alpha$ we get a standard bias-variance decomposition

$$\mathbf{E}\|\hat{\theta}_\alpha - \theta\|^2 = \sum_{k=1}^{n} \big[1 - H_\alpha(\lambda_k)\big]^2 \langle \theta, \psi_k \rangle^2 + \sigma^2 \sum_{k=1}^{n} \lambda_k H_\alpha^2(\lambda_k),$$

where $\langle \theta, \psi_k \rangle = \sum_{l=1}^{n} \theta(l)\psi_k(l)$.

Remarks:

- *The spectral regularization may improve substantially $\hat{\theta}_0$ when $\langle \theta, \psi_k \rangle^2$ are small for large k.*
- *The best regularization parameter depends on θ and therefore it should be data-driven.*

- Spectral cut-off (requires the SVD)

$$H_\alpha(\lambda) = \mathbf{1}\{\alpha\lambda \le 1\}.$$

- Tikhonov's regularization

$$\hat\theta_\alpha = \arg\min_\theta \Big\{ \|Y - A\theta\|^2 + \alpha\|\theta\|^2 \Big\}$$

or, equivalently,

$$\hat\theta_\alpha = [\alpha I + A^\top A]^{-1} A^\top Y, \quad H_\alpha(\lambda) = (1+\alpha\lambda)^{-1}.$$

- Landweber's iterations (solve $A^\top Y = A^\top A\theta$)

$$\hat\theta_i = \big[I - a^{-1}A^\top A\big]\hat\theta_{i-1} + a^{-1}A^\top Y.$$

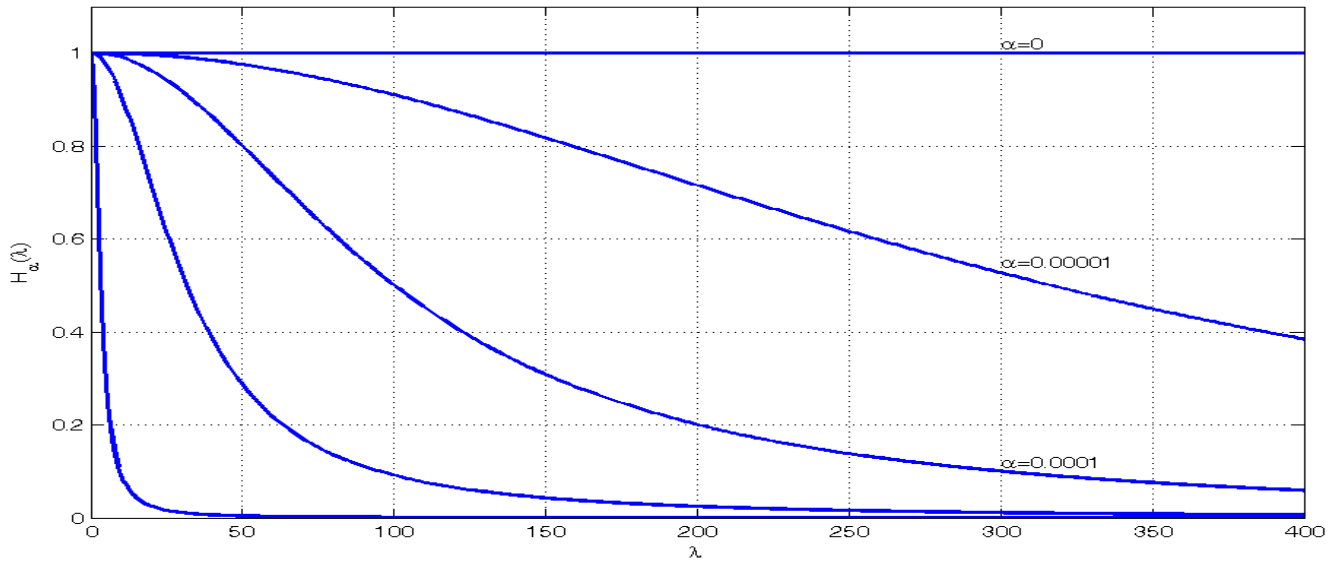
The iterations converge if $a\lambda_1 < 1$. It is easy to check that

$$H_\alpha(\lambda) = 1 - \big[1 - (a\lambda)^{-1}\big]^{1/\alpha}, \quad \alpha = 1/(i+1).$$

Ordered functions

In the above examples the families of functions (smoothers) $H_\alpha(\cdot),\ \alpha \in \mathbb{R}^+$ are *ordered* (see Kneip (1995))

- $0 \le H_\alpha(\lambda) \le 1$
- *for all* $\lambda \in \mathbb{R}^+$ $H_{\alpha_1}(\lambda) \ge H_{\alpha_2}(\lambda), \quad \alpha_1 \le \alpha_2.$



Our goal is to find the best estimate within the family spectral regularization methods

$$\hat{\theta}_\alpha = H_\alpha[(A^\top A)^{-1}](A^\top A)^{-1}A^\top Y,\ \alpha \in [0, \alpha^\circ].$$

In other words, we are looking for $\hat{\alpha}$ that minimizes

$$\mathbf{E}\|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \ \text{ uniformly in } \theta \in \mathbb{R}^n.$$

This idea puts into practice with the help of the empirical risk minimization principle :

$$\hat{\alpha} = \arg\min_\alpha R_\alpha[Y],\ \textit{where } R_\alpha[Y] = \|\hat{\theta}_0 - \hat{\theta}_\alpha\|^2 + \sigma^2 \mathit{Pen}(\alpha),$$

and $\mathit{Pen}(\alpha) : (0, \alpha^\circ] \to \mathbb{R}^+$ is a given function of α.

A good data-driven regularization should minimize in some sense the risk

$$L_\alpha(\theta) = \mathbf{E}\|\theta - \hat{\theta}_\alpha\|^2.$$

This is why, we are looking for a minimal penalty that ensures the following inequality

$$L_\alpha(\theta) \lesssim R_\alpha[Y] + \mathcal{C},$$

where $\mathcal{C}$ is a random variable that doesn't depend on α and θ. It is easy to check that

$$\mathcal{C} = -\|\theta - \hat{\theta}_0\|^2 = -\sigma^2 \sum_{k=1}^{n} \lambda_k \xi^2(k)$$

Traditional approach to solve this inequality is based on the unbiased risk estimation defining the penalty as a root of the equation

$$L_\alpha(\theta) = \mathbf{E}R_\alpha[Y] + \mathbf{E}\mathcal{C}.$$

Excess risk penalties

Unfortunately, thus obtained penalty is not good for ill-posed problems (see e.g. Cavalier and Golubev (2006)).
The main idea in this talk is to compute the penalty in a little bit different way, namely as a minimal root of the equation

$$\mathbf{E} \sup_{\alpha \le \alpha^\circ} \Big[L_\alpha(\theta) - R_\alpha[Y] - \mathcal{C} \Big]_+ \le K \mathbf{E} \Big[L_{\alpha^\circ}(\theta) - R_{\alpha^\circ}[Y] - \mathcal{C} \Big]_+ ,$$

where $[x]_+ = \max\{0, x\}$ and $K > 1$ is a constant.
Heuristic motivation: we are looking for the minimal penalty balancing the all excess risks.

It finds out that for ordered smoothers the penalty may be found as a solution of the marginal equation

$$\mathbf{E}\Big[L_\alpha(\theta) - R_\alpha[Y] - \mathcal{C}\Big]_+ \le \mathbf{E}\Big[L_{\alpha^\circ}(\theta) - R_{\alpha^\circ}[Y] - \mathcal{C}\Big]_+, \ \alpha \in [0, \alpha^\circ]$$

To compute the penalty, we assume that it has the following structure

$$Pen(\alpha) = 2\sum_{k=1}^{n} \lambda_k H_\alpha[\lambda_k] + (1+\gamma)Q(\alpha),$$

where $2\sum_{k=1}^{n} \lambda_k H_\alpha[\lambda_k]$ is the penalty related to the unbiased risk estimation. γ is a positive number and $Q(\alpha)$, $\alpha > 0$ is a positive function of α to be defined later on.

The large deviation approach results in the following algorithm for computing

$$Q(\alpha) = 2D(\alpha)\mu_\alpha \sum_{k=1}^{n} \frac{\rho_\alpha^2(k)}{1 - 2\mu_\alpha \rho_\alpha(k)},$$

where

$$D^2(\alpha) = 2\sum_{k=1}^{n} \lambda_k^2 \{2H_\alpha[\lambda_k] - H_\alpha^2[\lambda_k]\}^2,$$
$$\rho_\alpha(k) = \sqrt{2} D^{-1}(\alpha)\lambda_k \{2H_\alpha[\lambda_k] - H_\alpha^2[\lambda_k]\},$$

where μ_α is a root of equation

$$\sum_{k=1}^{n} F[\mu_\alpha \rho_\alpha(k)] = \log \frac{D(\alpha)}{D(\alpha^\circ)}, \ F(x) = \frac{1}{2}\log(1-2x) + x + \frac{2x^2}{1-2x}.$$

The following theorem provides the so-called oracle inequality which controls the performance of the method of the empirical risk minimization via the so-called penalized oracle risk defined by

$$r(\theta) \stackrel{\mathrm{def}}{=} \inf_{\alpha \le \alpha^\circ} \bar{R}_\alpha[\theta],$$

where

$$\bar{R}_\alpha[\theta] \stackrel{\mathrm{def}}{=} \mathbf{E}_\theta\{R_\alpha[Y] + \mathcal{C}\} = \mathbf{E}\|\theta - \hat{\theta}_\alpha\|^2 + (1+\gamma)\sigma^2 Q(\alpha).$$

Theorem

Uniformly in $\theta \in \mathbb{R}^n$,

$$\mathbf{E}_\theta\|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \le r(\theta)\left[1 + \frac{C}{\gamma^4}\log^{-1/2}\frac{Cr(\theta)}{\sigma^2\gamma D(\alpha^\circ)}\right].$$

This result represents a particular form of the so-called oracle inequality

$$\mathbf{E}_\theta \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \le r(\theta) + r(\theta)\Phi\left[\frac{\sigma^2 D(\alpha^\circ)}{r(\theta)}\right],$$

where $\Phi(\cdot)$ is a bounded function such that $\lim_{x\to 0}\Phi(x) = 0$. In other words, this inequality says that if the ratio $\sigma^2 D(\alpha^\circ)/r(\theta)$ is small then the risk of the method is close to the risk of the penalized oracle. On the other hand, if this ratio isn't small, then the risk of the method is of order of the oracle risk.
Note also that our oracle inequality holds whatever is the ill-posedness of the underlying inverse problem. What depends on the ill-posedness is solely the extra penalty $(1+\gamma)\sigma^2 Q(\alpha)$.

For $Q(\alpha)$ we have the following bounds

$$D(\alpha)\sqrt{\log[D(\alpha)/D(\alpha^\circ)]} \le Q(\alpha) \le CD(\alpha)\log[D(\alpha)/D(\alpha^\circ)].$$

Therefore, if the inverse problem is not severely ill-posed, i.e. $\lambda(k) \le Ck^\beta$, then for small α

$$\sum_{k=1}^{n} \lambda_k H_\alpha^2[\lambda_k] \gg Q(\alpha).$$

So, the risk of penalized oracle is close to the risk of the ideal oracle $\inf_{\alpha \le \alpha^\circ} \mathbf{E}\|\theta - \hat{\theta}_\alpha\|^2$.

On the other hand, if the inverse problem is severely ill-posed, i.e. $\lambda(k) \approx \exp(\beta k)$, then

$$\sum_{k=1}^{n} \lambda_k H_\alpha^2[\lambda_k] \ll Q(\alpha)$$

and the risk of penalized oracle is essentially greater than that one of the ideal oracle. However, neither this upper bound nor the extra penalty can be improved.

Now we consider the case where σ is unknown. To chose α in this situation, we plug-in a standard estimator for σ^2 in the penalized empirical risk, thus arriving at the following formula for the empirical risk

$$R_\alpha^\sigma[Y] \stackrel{\text{def}}{=} \|\hat{\theta}_0 - \hat{\theta}_\alpha\|^2 + \frac{\|Y - A\hat{\theta}_\alpha\|^2}{\|1 - H_\alpha\|^2} Pen(\alpha).$$

Finally, we compute the data-driven regularization parameter as follows:

$$\hat{\alpha} = \underset{\alpha_\circ \le \alpha \le \alpha^\circ}{\arg\min}\, R_\alpha^\sigma[Y].$$

The following theorem controls the performance of the method of the empirical risk minimization via the penalized oracle risk defined by

$$r(\theta) \stackrel{\mathrm{def}}{=} \inf_{\alpha_\circ \le \alpha \le \alpha^\circ} \bar{R}^{\sigma}_{\alpha}[\theta],$$

where

$$\bar{R}^{\sigma}_{\alpha}[\theta] \stackrel{\mathrm{def}}{=} \mathbf{E}_\theta\{R_\alpha[Y] + \mathcal{C}\} = \mathbf{E}\|\theta - \hat{\theta}_\alpha\|^2 + (1+\gamma)\sigma^2 Q(\alpha) + \frac{Pen(\alpha)}{\|1 - H_\alpha\|^2} \sum_{k=1}^{n} \{1 - H_\alpha[\lambda(k)]\}^2 \frac{\theta^2(k)}{\lambda(k)}.$$

Denote also for brevity

$$\|H_\alpha\|_\lambda^2 = \sum_{k=1}^{n} \lambda(k) H_\alpha^2[\lambda(k)],$$

$$\Psi(x) = x \log^2(\exp(1) + x),$$

$$\Sigma_\alpha = \|1 - H_\alpha\| \sqrt{2 \log\log \frac{\|1 - H_{\alpha^\circ}\| \exp(2)}{\|1 - H_\alpha\|}}.$$

$$q = \max_{\alpha \in [\alpha_\circ, \alpha^\circ]} \frac{47 Pen(\alpha) \Sigma_\alpha \log\left[Q^\circ(\alpha) + \|H_\alpha\|_\lambda^2\right]}{\Sigma_{\alpha^\circ} \|1 - H_\alpha\|^2 [Q^\circ(\alpha) + \|H_\alpha\|_\lambda^2]}$$

$$\approx \sqrt{\frac{\log\log(n)}{n}} \max_{\alpha \in [\alpha_\circ, \alpha^\circ]} \frac{Pen(\alpha) \log[\|H_\alpha\|_\lambda^2 + Q(\alpha)]}{[\|H_\alpha\|_\lambda^2 + Q(\alpha)]}.$$

Theorem

Uniformly in $\theta \in \mathbb{R}^n$,

$$\mathbf{E}_\theta \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \leq [1 + C\Psi(q)] r(\theta) + \frac{Cr(\theta)}{[1 - C\Psi(q)]\gamma^4} \log^{-1/2} \frac{Cr(\theta)}{\sigma^2 \gamma D(\alpha^\circ)}.$$

There are two main distinctions with respect to the case where the noise variance is known. The first one is that in the penalized oracle risk there is an additional term, namely

$$\frac{Pen(\alpha)}{\|1 - H_\alpha\|^2} \sum_{k=1}^{n} \{1 - H_\alpha[\lambda(k)]\}^2 \frac{\theta^2(k)}{\lambda(k)}.$$

Since

$$\sum_{k=1}^{n} \{1 - H_\alpha[\lambda(k)]\}^2 \frac{\theta^2(k)}{\lambda(k)} \le \mathbf{E}\|\theta - \hat{\theta}_\alpha\|^2$$

and we may chose $\alpha_\circ$ so that $\|1 - H_\alpha\|^2 \ge Cn$ and $Pen(\alpha) \ll n$ for all $\alpha \ge \alpha_\circ$, this term is typically small.

The second distinction is related to the parameter

$$q \gtrsim \sqrt{\frac{\log\log(n)}{n}} \max_{\alpha\in[\alpha_\circ,\alpha^\circ]} \frac{\log\left[Q(\alpha)\right]\sum_{k=1}^n \lambda_k H_\alpha(\lambda_k)}{\left[\sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) + Q(\alpha)\right]}$$

which is typically small but for some regularization methods it may be large. Indeed, for Tikhonov's regularization with $\lambda(k) \asymp k^\beta$, $\beta > 1$, we have

$$\sum_{k=1}^n \lambda_k H_\alpha(\lambda_k) \asymp \frac{n}{\alpha}, \;\; Q(\alpha) \approx \frac{\sqrt{n}}{\alpha}\log\frac{\sqrt{n}}{\alpha}.$$

So, $q \asymp C\sqrt{\log\log(n)}$, thus demonstrating that the oracle inequality blows up.