The law of the iterated logarithm and data-driven regularizations of ill-posed inverse problems

Yuri Golubev

CNRS, Université de Provence and Institute for Information Transmission Problems

SAPS-8, Le Mans, 21–24 March 2011

Outline of the talk

1. Spectral regularization
   - Ordered regularizations
2. Empirical Risk Minimization
   - Excess risk penalties
3. Oracle inequalities and the law of the iterated logarithm

This talk focuses on recovering an unknown function $\theta(\cdot) \in L_2[0,1]$ from the noisy data

$$dY(t) = \left(\int_0^1 A(t,u)\theta(u)\,du \right) dt + \sigma dW(t), \;\; t \in [0,1],$$

where

- $A(\cdot,\cdot) \in L_2[0,1] \times L_2[0,1]$ is a known kernel
- $W(\cdot)$ is a standard Wiener process
- σ is a known noise level.

Maximum likelihood estimator

For brevity, denote by $\mathbb{A}$ the compact operator $L_2[0,1] \to L_2[0,1]$

$$\mathbb{A}x = \int_0^1 A(t,u)x(u)\,du$$

We begin with ML estimator

$$\hat{\theta}_0 = \arg\min_{\theta}\Big\{ \|\mathbb{A}\theta\|^2 - 2\langle \dot{Y}, \mathbb{A}\theta\rangle \Big\}.$$

With a simple algebra we obtain

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

Since the operator $(\mathbb{A}^\top\mathbb{A})^{-1}\mathbb{A}^\top$ is unbounded, $\mathbf{E}\|\hat{\theta}_0 - \theta\|^2 = \infty$.

The Tikhonov-Phillips regularization is computed as follows

$$\hat{\theta}_\alpha = \arg\min_\theta \Big\{ \|\mathbb{A}\theta\|^2 - 2\langle \dot{Y}, \mathbb{A}\theta\rangle + \alpha\|\theta\|^2 \Big\},$$

where $\alpha > 0$ is a regularization parameter. It is easy to check that

$$\hat{\theta}_\alpha = (\mathbb{A}^\top \mathbb{A} + \alpha I)^{-1} \mathbb{A}^\top \dot{Y}.$$

The simplest way to understand how does this method work, is based on the SVD.

Let λ_k and $\phi_k(\cdot)$ be eigenvalues and eigenfunctions of $\mathbb{A}\mathbb{A}^\top$, i.e.

$$\mathbb{A}\mathbb{A}^\top \phi_k = \lambda_k \phi_k, \ k = 1, 2, \ldots$$

Then we get the equivalent spectral decomposition of our data

$$y_k \stackrel{\text{def}}{=} \langle \mathbb{A}^\top Y, \phi_k \rangle = \lambda_k \langle \theta, \phi_k \rangle + \sigma \sqrt{\lambda_k} \xi_k, \quad k = 1, 2, \ldots,$$

where ξ_k are i.i.d. $\mathcal{N}(0, 1)$. So, we get

$$\langle \hat{\theta}_0, \phi_k \rangle = \langle \theta, \phi_k \rangle + \sigma \lambda_k^{-1/2} \xi_k$$

and

$$\langle \hat{\theta}_\alpha, \phi_k \rangle = \frac{\lambda_k}{\alpha + \lambda_k} \langle \hat{\theta}_0, \phi_k \rangle$$

The mean square risk of the Tikhonov-Phillips regularization is computed as follows :

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

where $h_\alpha(\lambda) = \lambda/(\alpha + \lambda)$.
Note that

- the risk is bounded if $\sum_{k=1}^{\infty}\lambda_k < \infty$,
- the risk may be improved with a properly chosen α.

Spectral regularizations

The basic idea in the spectral regularization is to smooth $\hat{\theta}_0$ with the help of a family of linear operators $\mathbb{H}_\alpha, \alpha \in \mathbb{R}^+$

$$\hat{\theta}_\alpha = \mathbb{H}_\alpha \hat{\theta}_0 = H_\alpha\big[(\mathbb{A}^\top \mathbb{A})^{-1}\big](\mathbb{A}^\top \mathbb{A})^{-1}\mathbb{A}^\top \dot{Y},$$

where $H_\alpha(\lambda),\ \lambda \in \mathbb{R}^+$ is an analytical function such that

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

Bias-variance decomposition

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

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

Remark:

- *The best regularization parameter depends on θ and therefore it should be data-driven.*

Ordered regularizations

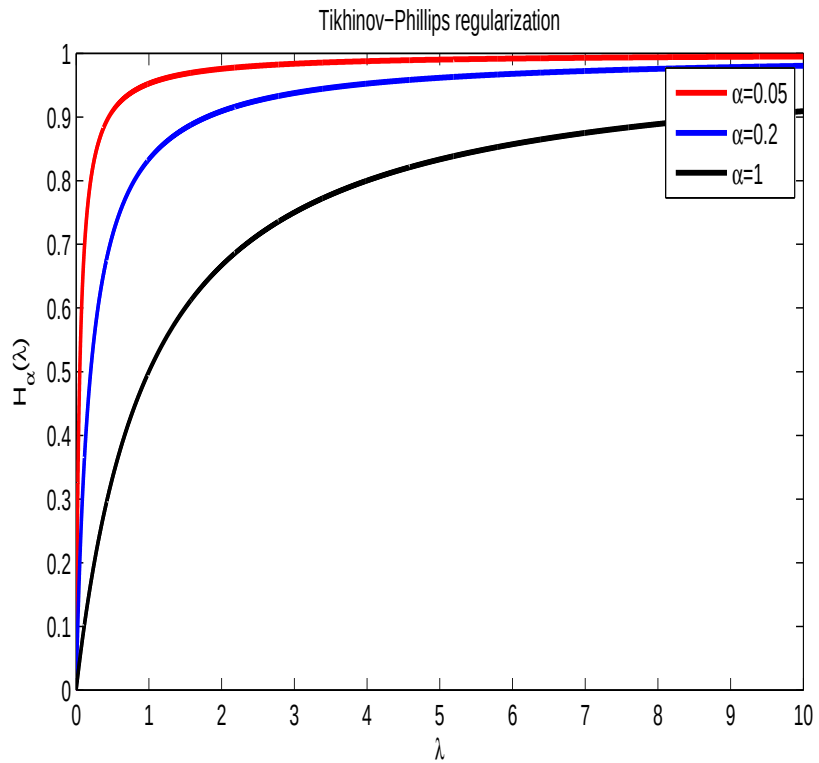
In what follows we shall deal with ordered families of functions $H_\alpha(\cdot),\ \alpha \in \mathbb{R}^+$ (Kneip (1995)) :

- $0 \le H_\alpha(\lambda) \le 1$
- *if some* $\alpha_1, \alpha_2, \lambda_0 \in \mathbb{R}^+$

$$H_{\alpha_1}(\lambda_0) > H_{\alpha_2}(\lambda_0),$$

then for all $\lambda \in \mathbb{R}^+$

$$H_{\alpha_1}(\lambda) \ge H_{\alpha_2}(\lambda).$$



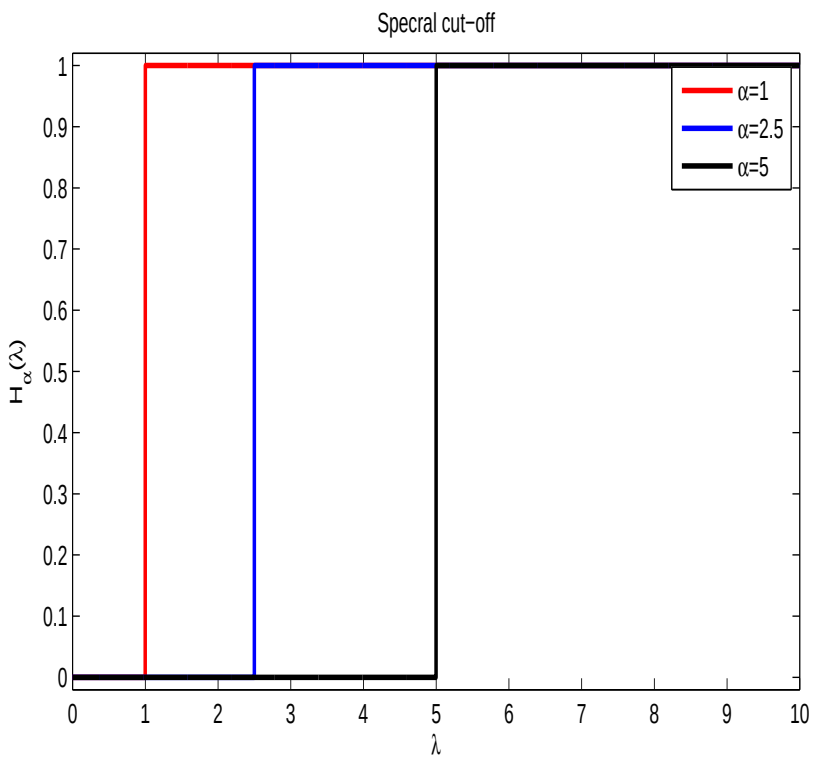


Figure: $H_\alpha(\lambda) = \lambda/(\alpha+\lambda)$

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

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

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

In other words, we are looking for $\hat{\alpha} \in [0, \alpha^\circ]$ that minimizes

$$\mathbf{E}\|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 = \mathbf{E}\big\{\|\hat{\theta}_{\hat{\alpha}}\|^2 - 2\langle\theta, \hat{\theta}_{\hat{\alpha}}\rangle + \|\theta\|^2\big\}.$$

This idea may be put into practice with the help of the empirical risk minimization principle :

$$\hat{\alpha} = \arg\min_\alpha R_\alpha[Y, Pen],$$

where

$$R_\alpha[Y, Pen] = \|\hat{\theta}_\alpha\|^2 - 2\langle\hat{\theta}_0, \hat{\theta}_\alpha\rangle + \sigma^2 Pen(\alpha),$$

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

Since good data-driven regularizations should minimize in some sense the risk

$$L_\alpha[\theta] = \mathbf{E}\big\{\|\hat{\theta}_\alpha\|^2 - 2\langle\theta, \hat{\theta}_\alpha\rangle\big\},$$

heuristically, we are looking for a minimal penalty that ensures the following inequality

$$L_\alpha[\theta] \lesssim R_\alpha[Y, Pen].$$

The traditional approach in solving this inequality is based on the unbiased risk estimation (Akaike (1973)). This method prompts to compute the penalty as a root of the equation

$$L_\alpha[\theta] = \mathbf{E}R_\alpha[Y, Pen].$$

With a simple algebra we get

$$Pen(\alpha) = Pen_0(\alpha) \stackrel{\text{def}}{=} 2\sum_{k=1}^{\infty} \lambda_k^{-1} H_\alpha(\lambda_k).$$

Excess risk penalties

Unfortunately, thus obtained penalty is not good for ill-posed problems.
The main idea in this talk is to compute the penalty as a minimal root of the equation

$$\mathbf{E} \sup_{\alpha \le \alpha^\circ} \Big\{ L_\alpha[\theta] - R_\alpha[Y, Pen] \Big\}_+ \le K \mathbf{E} \Big\{ L_{\alpha^\circ}[\theta] - R_{\alpha^\circ}[Y, Pen_0] \Big\}_+ ,$$

where $\{x\}_+ = \max\{0, x\}$ and $K > 1$ is a constant.
Heuristic motivation: we are looking for the minimal penalty such that for any (data-driven) $\alpha', \alpha'' \in [0, \alpha^\circ]$

$$\Big\{ L_{\alpha'}[\theta] - R_{\alpha'}[Y, Pen] \Big\}_+ \approx \Big\{ L_{\alpha''}[\theta] - R_{\alpha''}[Y, Pen] \Big\}_+ .$$

In order to compute a nearly optimal penalty, consider the following random process

$$\begin{aligned}\zeta_\alpha =& L_\alpha[\theta] - R_\alpha[Y, Pen_0] \\ & -2\sigma \sum_{k=1}^{\infty} \frac{2H_\alpha(\lambda_k) - H_\alpha^2(\lambda_k)}{\sqrt{\lambda_k}} \langle \theta, \phi_k \rangle \xi_k \\ =& \sigma^2 \sum_{k=1}^{\infty} \frac{2H_\alpha(\lambda_k) - H_\alpha^2(\lambda_k)}{\lambda_k} (\xi_k^2 - 1).\end{aligned}$$

Our first step is to compute the minimal penalty such that

$$\mathbf{E} \sup_{\alpha \le \alpha^\circ} \Big\{ \zeta_\alpha - \sigma^2 Pen(\alpha) \Big\}_+ = K \mathbf{E} \Big\{ \zeta_{\alpha^\circ} \Big\}_+ .$$

Let $Pen_1(\alpha)$ be a root of the equation

$$\mathbf{E} \Big\{ \zeta_\alpha - \sigma^2 Pen_1(\alpha) \Big\}_+ = \mathbf{E} \Big\{ \zeta_{\alpha^\circ} \Big\}_+ .$$

To compute the upper bound for $Pen_1(\alpha)$, we make use of the large deviation principle. Let

$$D^2(\alpha) \stackrel{\text{def}}{=} \sigma^{-2}\mathbf{E}\zeta_\alpha^2.$$

Then

$$Pen_1(\alpha) \le 2D(\alpha)\mu_\alpha \sum_{k=1}^{n} \frac{\rho_\alpha^2(k)}{1-2\mu_\alpha\rho_\alpha(k)},$$

where

$$\rho_\alpha(k) \stackrel{\text{def}}{=} \sqrt{2}\frac{2H_\alpha(\lambda_k) - H_\alpha^2(\lambda_k)}{\lambda_k D(\alpha)},$$

and μ_α is a root of equation

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

Theorem

For any $\gamma > 0$

$$\mathbf{E} \sup_{\alpha \le \alpha^\circ} \left\{ \zeta_\alpha - \sigma^2 Pen_\gamma(\alpha) \right\}_+ \le \frac{C\sigma^2 D(\alpha^\circ)}{\gamma},$$

where

$$Pen_\gamma(\alpha) = Pen_1(\alpha) + (1+\gamma)\sqrt{2} D(\alpha) \log\left[\frac{1}{\gamma} \log \frac{D(\alpha)}{D(\alpha^\circ)} \right].$$

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

$$r(\theta) \stackrel{\text{def}}{=} \inf_{\alpha \le \alpha^\circ} \left\{ \mathbf{E}\|\theta - \hat{\theta}_\alpha\|^2 + \sigma^2 Pen_1(\alpha) \right\}.$$

Theorem

Let

$$\hat{\alpha}_\gamma = \arg\min_\alpha \Big\{ \|\hat{\theta}_\alpha\|^2 - 2\langle \hat{\theta}_0, \hat{\theta}_\alpha \rangle + \sigma^2 Pen_0(\alpha) + \sigma^2 Pen_\gamma(\alpha) \Big\}.$$

with

$$Pen_\gamma(\alpha) = Pen_1(\alpha) + (1+\gamma)\sqrt{2} D(\alpha) \log \left[\frac{1}{\gamma} \log \frac{D(\alpha)}{D(\alpha^\circ)} \right].$$

Then uniformly in $\theta \in \mathbb{R}^n$,

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

where $\Phi(\cdot)$ *is a bounded function such that* $\lim_{x \to 0} \Phi(x) = 0$.

In other words,

- 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

$$\mathbf{E}_\theta \|\theta - \hat{\theta}_{\hat{\alpha}_\gamma}\|^2 \approx r(\theta),$$

- if this ratio isn't small, then the risk of the method is of order of the oracle risk

$$\mathbf{E}_\theta \|\theta - \hat{\theta}_{\hat{\alpha}_\gamma}\|^2 \approx Cr(\theta), \ C > 1.$$

Note also that the oracle inequality holds whatever is the ill-posedness of the underlying inverse problem. What depends on the ill-posedness is solely the extra penalty $Pen_1(\alpha)$.

For $Pen_1(\alpha)$ we have the following bounds

$$D(\alpha)\sqrt{\log[D(\alpha)/D(\alpha^\circ)]} \le Pen_1(\alpha) \le CD(\alpha)\log[D(\alpha)/D(\alpha^\circ)].$$

Therefore, if the inverse problem is not severely ill-posed, i.e. $\lambda(k) \ge Ck^{-\beta}$ $\beta \in [0,\infty)$, then typically for small α

$$\mathbf{Var}(\theta - \hat{\theta}_\alpha) = \sigma^2 \sum_{k=1}^{n} \lambda_k H_\alpha^2(\lambda_k) \gg \sigma^2 Pen_1(\alpha).$$

So, the risk of penalized oracle is close to the risk of the ideal oracle

$$r(\theta) = \inf_{\alpha \le \alpha^\circ} \left\{ \mathbf{E}\|\theta - \hat{\theta}_\alpha\|^2 + \sigma^2 Pen_1(\alpha) \right\} \approx \inf_{\alpha \le \alpha^\circ} \mathbf{E}\|\theta - \hat{\theta}_\alpha\|^2$$

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

$$\mathbf{Var}(\theta - \hat{\theta}_\alpha) \ll Pen_1(\alpha)$$

and

$$r(\theta) \gg \inf_{\alpha \le \alpha^\circ} \mathbf{E}\|\theta - \hat{\theta}_\alpha\|^2$$

but the upper bound

$$\inf_{\hat{\alpha}} \mathbf{E}\|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \le r(\theta)$$

cannot be improved.