Linear Systems and Inverse problems

Ulysse Marteau-Ferey

Learning problem and regularized empirical risk minimization

- Input: random variable $X \in \mathcal{X}$;
- Output: random variable $Y \in \mathcal{Y}$;
- Law of (X, Y) : $\rho(x, y)$
- Objective : find a predictor $f : \mathcal{X} \to \mathcal{Y}$.
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Learning problem

$$\inf_{f \in \mathcal{F}(\mathcal{X},\mathcal{Y})} L(f) := \mathbb{E}\left[\ell(f(x), y)\right] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y)\, d\rho(x, y)$$

Classical way of tackling the learning problem

Learning problem

$$\inf_{f \in \mathcal{F}(\mathcal{X},\mathcal{Y})} L(f) := \mathbb{E}\left[\ell(f(x), y)\right] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y)\, d\rho(x, y)$$

Access to ρ only through $\mathbf{z} = (z_i)_{1 \leqslant i \leqslant n}$ where $z_i = (x_i, y_i)$ are training samples

Regularized empirical risk minimization

$$\inf_{f \in \mathcal{H}} \widehat{L}_\lambda(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) + \frac{\lambda}{2}\Omega(f)$$

- $\mathcal{H}$ is a space of functions;
- Ω is a regularizer.

Least squares case

- We take $\ell(y, y') = \frac{1}{2}\|y - y'\|^2$;
- we assume that $Y \in L^2(\mathcal{Y}, \rho_Y)$.

The minimizer exists

$$g_* = \mathbb{E}\left[Y|X\right], \; g_*(x) = \int_{\mathcal{Y}} y \; d\rho(y|x), \;\; \in L^2(\mathcal{X}, \rho_X)$$

Consequence: one needs only to solve

$$\inf_{f \in L^2(\mathcal{X}, \rho_X)} \mathbb{E}\left[\|f(X) - Y\|^2\right]$$

ERM

Regularized empirical risk minimization

$$\inf_{f \in \mathcal{H}} \widehat{L}_\lambda(f) := \frac{1}{n} \sum_{i=1}^{n} \|f(x_i) - y_i\|^2 + \lambda \Omega(f)$$

Question

What space $\mathcal{H}$ of functions and what penalty Ω ?

Requirements:

- $\mathcal{H} \hookrightarrow L^2(\mathcal{X}, \rho_X)$
- Solvable : finite dimensional ?

Classical parameterized version :

$$\mathcal{H} = \{f_\theta \ : \ \theta \in \Theta\}, \ \Theta \subset \mathbb{R}^d$$

Alternative ?

RKHS

RKHS $\mathcal{H}_K$

- $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ measurable
- K positive definite, i.e.

$$\forall n \in \mathbb{N},\ \mathbf{x} = (x_i)_{1 \leqslant i \leqslant n},\ \alpha = (\alpha_i)_{1 \leqslant i \leqslant n},\ \sum_{1 \leqslant i,j \leqslant n} \alpha_i \alpha_j K(x_i, x_j) \geqslant 0$$

- $\forall x \in \mathcal{X},\ K_x := K(x, \,.\,) \in \mathcal{H}_K$
- $\mathcal{H}_K := \overline{\text{Span}\left(\{K_x,\ x \in \mathcal{X}\}\right)}$ endowed with $\langle \,.\,,\,.\, \rangle_{\mathcal{H}}$ so that $\langle K_x, K_{x'} \rangle_{\mathcal{H}} = K(x, x')$

Properties/Assumptions

- $\forall f \in \mathcal{H},\ f(x) = \langle f, K_x \rangle_{\mathcal{H}}$
- Assume $K(x,x) \leqslant \kappa^2$ such that

$$\forall f \in \mathcal{H},\ \int_{\mathcal{X}} \|f(x)\|^2 d\rho_X(x) = \int_{\mathcal{X}} \|f \cdot K_x\|^2 d\rho_X(x) \leqslant \kappa^2 \|f\|^2_{\mathcal{H}}$$

Formulation of the problem

Regularized empirical risk minimization

$$\widehat{f}_\lambda = \arg\min_{f\in\mathcal{H}} \widehat{L}_\lambda(f) := \frac{1}{n}\sum_{i=1}^{n} \underbrace{\|f(x_i)-y_i\|^2}_{\|f\cdot K_{x_i}-y_i\|^2} + \lambda\|f\|^2_{\mathcal{H}}$$

Representer theorem :

$$\widehat{f}_\lambda = \sum_{i=1}^{n} \alpha_i \, K_{x_i}$$

Equivalent problem, linear and finite dimensional !

$$\inf_{\alpha\in\mathbb{R}^n} \alpha^T(K_{nn}^2 + \lambda n K_{nn})\alpha - 2(K_{nn}\mathbf{y})\cdot\alpha \quad \Leftrightarrow \quad (K_{nn}+\lambda n\mathbf{I})\alpha = \mathbf{y}$$

Linear inverse problem

Inverse problem setting

- $\mathcal{H}, \mathcal{K}$ two Hilbert spaces
- An operator $\mathbf{A} \in \mathcal{L}(\mathcal{H}, \mathcal{K})$
- An objective $g \in \mathcal{K}$.

The aim is to reconstruct a solution to

$$\mathbf{A}f = g$$

Machine Learning problem

Ideal problem

$f(X) = Y, \ f \in \mathcal{H}$ which can be reformulated as $Sf = Y$.

- $S : \mathcal{H} \to L^2(\mathcal{X} \times \mathcal{Y}, \rho)$ such that $(Sf)(x, y) = f \cdot K_x = f(x)$
- $S^* : L^2(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}$, such that $S^* g = \int_{\mathcal{X} \times \mathcal{Y}} g(x, y) K_x \, d\rho(x, y)$
- $C = S^* S : \mathcal{H} \to \mathcal{H}$ such that $\langle f, Cf \rangle_{\mathcal{H}} = \int_{\mathcal{X}} \|f(x)\|^2 d\rho_X(x)$

Ill-posed problems

Possibly no solution to $\mathbf{A}f = g$ (or obviously $Sf = Y$)!

Alternative problem

$$\inf_{f \in \mathcal{H}} \|\mathbf{A}f - g\|_{\mathcal{K}}^2, \qquad \inf_{f \in \mathcal{H}} \|Sf - Y\|_{L^2(\mathcal{X}\times\mathcal{Y},\rho)}^2$$

$\mathbf{P}_{\mathbf{A}}$ orthogonal projector on $\overline{\mathrm{range}(\mathbf{A})}$

$$\|\mathbf{A}f - g\|_{\mathcal{K}}^2 = \|\mathbf{A}f - \mathbf{P}_{\mathbf{A}}g\|_{\mathcal{K}}^2 + \underbrace{\|(\mathbf{I} - \mathbf{P}_{\mathbf{A}})g\|_{\mathcal{K}}^2}_{\text{inevitable error}}$$

$$\mathrm{range}(S) \subset L^2(\mathcal{X}, \rho_X) \implies \mathbf{P}_S Y = \mathbf{P}_S g_*,\ g_*(x) = \int_{\mathcal{Y}} y \, d\rho(y|x)$$

Solution if it exists :

$$\mathbf{A}^*\mathbf{A}\, f = \mathbf{A}^* g, \qquad C\, f = S^* Y = S^* \mathbf{P}_S g_*$$

Approximations

Inverse problem setting

Let $\delta = (\delta_1, \delta_2) \in \mathbb{R}_+^2$.

- An approximation space $\widetilde{\mathcal{K}}$
- An approximation of $\mathbf{A}$: $\mathbf{A}_{\delta_1} : \mathcal{H} \to \widetilde{\mathcal{K}}$ such that $\|\mathbf{A}^*\mathbf{A} - \mathbf{A}_{\delta_1}^*\mathbf{A}_{\delta_1}\| \leqslant \delta_1$
- An approximation of g : $g_{\delta_2} \in \widetilde{\mathcal{K}}$ such that $\|\mathbf{A}_{\delta_1}^* g_{\delta_2} - \mathbf{A}^* g\| \leqslant \delta_2$

Machine Learning : approximation space limited by the data

- An approximation space $\mathbb{R}^n$
- An approximation of Y : $\mathbf{y} = \frac{1}{\sqrt{n}}(y_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$
- An approximation of S : $S_n : \mathcal{H} \to \mathbb{R}^n$,
  $S_n f = \frac{1}{\sqrt{n}}(f \cdot K_{x_i})_{1 \leqslant i \leqslant n} = \frac{1}{\sqrt{n}}(f(x_i))_{1 \leqslant i \leqslant n}$.

Naive method

Potential solutions:

$$\mathbf{A}^* \mathbf{A}\, f = \mathbf{A}^* g, \qquad S^* S\, f = S^* Y$$

$$\mathbf{A}_{\delta_1}^* \mathbf{A}_{\delta_1} f_\delta = \mathbf{A}_{\delta_1}^* g_{\delta_2}, \qquad S_n^* S_n \widehat{f} = S_n^* \mathbf{y}$$

Define $C_n = S_n^* S_n : \mathcal{H} \to \mathcal{H}$ such that $C_n = \sum_{i=1}^n K_{x_i} \otimes K_{x_i}$

Equivalence with empirical risk minimization for learning

$$C_n \widehat{f} = S_n^* \mathbf{y} \Leftrightarrow \widehat{f} \in \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \|f(x_i) - y_i\|^2$$

Problems :

- The second problem is not necessarily solvable (it is in finite dimension)
- We want the reconstruction error to be small, i.e.

$$\|Af_\delta - \mathbf{P}_\mathbf{A} g\|^2 \xrightarrow[\delta \to 0]{} 0$$

- Generalization error

Tikhonov regularization

Approximations : $\|\mathbf{A}^*\mathbf{A} - \mathbf{A}_{\delta_1}^*\mathbf{A}_{\delta_1}\| \leqslant \delta_1, \|\mathbf{A}_{\delta_1}^* g_{\delta_2} - \mathbf{A}^* g\| \leqslant \delta_2$

Regularization method

For a given $\lambda > 0$, choose

$$f_{\lambda,\delta} = \left(\mathbf{A}_{\delta_1}^*\mathbf{A}_{\delta_1} + \lambda \mathbf{I}\right)^{-1} \mathbf{A}_{\delta_1}^* g_{\delta_2} = \arg\min_{f \in \mathcal{H}} \|\mathbf{A}_\delta f - g_{\delta_2}\|^2 + \lambda \|f\|^2$$

Aim : choose $\lambda(\delta)$ well to have

$$\|Af_{\lambda(\delta),\delta} - \mathbf{P}_{\mathbf{A}} g\|^2 \xrightarrow[\delta \to 0]{} 0$$

Analysis of the error for Tikhonov regularization

Define $f_\lambda = (\mathbf{A}^*\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^* g = (\mathbf{A}^*\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^*\mathbf{P}_\mathbf{A} g = \arg\min_{\|Af-g\|^2+\lambda\|f\|^2}$.

Decomposition

If $\lambda \geqslant 2\delta_1$, then the following holds:

$$\|\mathbf{A}f_{\lambda,\delta} - \mathbf{P}_\mathbf{A} g\| \leqslant \underbrace{\|\mathbf{A}f_\lambda - \mathbf{P}_\mathbf{A} g\|}_{\mathcal{S}(\lambda)} + \frac{\delta_1}{\lambda} + \frac{\delta_2}{\sqrt{\lambda}}.$$

$\mathcal{S}(\lambda)$ caracterizes the between $P_\mathbf{A} g$ and the range of $\mathbf{A}$.

If $\mathbf{P}_\mathbf{A} g \in \mathrm{range}(\mathbf{A})$, $\mathcal{S}(\lambda) \leqslant C\lambda^{1/2}$

Equivalent to solving the regularized ERM

Equivalence

$$\widehat{f}_\lambda = (S_n^* S_n + \lambda I)^{-1} S_n^* \mathbf{y} \Leftrightarrow \widehat{f}_\lambda \in \arg\min_{f \in \mathcal{H}} \|S_n f - \mathbf{y}\|_{\mathbb{R}^n}^2 + \lambda \|f\|_{\mathcal{H}}^2$$

$$S_n f = \frac{1}{\sqrt{n}} (f(x_i))_{1 \leqslant i \leqslant n} \quad \Longrightarrow \quad \widehat{f}_\lambda \in \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \|f(x_i) - y_i\|^2 + \lambda \|f\|_{\mathcal{H}}^2$$

Tikhonov regularization amounts to solving the reguralized ERM problem

$$\delta_2 = \|S_n^* \mathbf{y} - S^* y\| = \|\tfrac{1}{n} \textstyle\sum_{i=1}^n y_i K_{x_i} - \mathbb{E}\left[Y K_X\right]\|$$

$$\delta_1 = \|C_n - C\| = \|\tfrac{1}{n} \textstyle\sum_{i=1}^n K_{x_i} \otimes K_{x_i} - \mathbb{E}\left[K_X \otimes K_X\right]\|$$

Bounds in high probability

Very naive bounds: with probability at least $1-\delta$

$$\delta_1 \leqslant \frac{\kappa^2 \log \frac{1}{\delta}}{\sqrt{n}}, \quad \delta_2 \leqslant \frac{\|y\|_\infty \kappa \log \frac{1}{\delta}}{\sqrt{n}}$$

Statistical bound on ERM

With probability at least $1-\delta$,

$$\|\widehat{f}_\lambda(x) - \mathbf{P}_S g_*(x)\|_{L^2(\mathcal{X},\rho_X)} \leqslant \mathcal{S}(\lambda) + \left(\frac{\kappa^2}{\sqrt{n}\lambda} + \frac{\kappa\|y\|_\infty}{\sqrt{\lambda n}}\right).$$

Refinement of the bound

Consider $\mathbf{A}_\delta^*\mathbf{A}_\delta \approx \mathbf{A}^*\mathbf{A}$, $\mathbf{A}_\delta^* g_\delta \approx \mathbf{A}^* g$. Define

$$\epsilon_1 = \|(\mathbf{A}^*\mathbf{A} + \lambda\mathbf{I})^{-1/2}(\mathbf{A}_\delta^*\mathbf{A}_\delta - \mathbf{A}^*\mathbf{A})(\mathbf{A}^*\mathbf{A} + \lambda\mathbf{I})^{-1/2}\|$$

$$\epsilon_2 = \|(\mathbf{A}^*\mathbf{A} + \lambda\mathbf{I})^{-1/2}(\mathbf{A}_\delta^* g_\delta - \mathbf{A}^* g)\|$$

$$f_{\lambda,\delta} = \arg\min_f \|\mathbf{A}_\delta f - g_\delta\|^2 + \lambda\|f\|^2$$

Tikhonov, refinement

If $\epsilon_1 < \frac{1}{2}$,

$$\|Af_{\lambda,\delta} - \mathbf{P}_\mathbf{A} g\| \leqslant \mathcal{S}(\lambda) + \epsilon_1 + \epsilon_2$$

$\mathcal{S}(\lambda) = \|Af_\lambda - \mathbf{P}_\mathbf{A} g\|$

$$\begin{aligned}\epsilon_1 &= \|(C_n + \lambda \mathbf{I})^{-1/2}(C_n - C)(C_n + \lambda \mathbf{I})^{-1/2}\| \\ &= \|\frac{1}{n}\sum_{i=1}^{n} C_\lambda^{-1/2} K_{x_i} \otimes K_{x_i} C_\lambda^{-1/2} - \mathbb{E}\left[C_\lambda^{-1/2} K_X \otimes K_X C_\lambda^{-1/2}\right]\|.\end{aligned}$$

$$\epsilon_2 = \|(C + \lambda \mathbf{I})^{-1/2}(S_n^* \mathbf{y} - S^* y)\| = \|\frac{1}{n}\sum_{i=1}^{n} C_\lambda^{-1/2} y_i K_{x_i} - \mathbb{E}\left[C_\lambda^{-1/2} Y K_X\right]\|$$

High probability results

Let $d_\infty(\lambda) = \sup_{x \in supp \rho_X} \|C_\lambda^{-1/2} K_x\|^2$. Then with probability at least $1 - \delta$, for any $\lambda \leqslant \|C\|$, $\epsilon_1 \leqslant \sqrt{\frac{d_\infty(\lambda)}{n}} \log \frac{1}{\delta}$ and $\epsilon_2 = \|y\|_\infty \sqrt{\frac{d_\infty(\lambda)}{n}} \log \frac{1}{\delta}$

Final bound

$$\|\widehat{f}_\lambda(x) - \mathbf{P}_S g_*(x)\|_{L^2(\mathcal{X}, \rho_X)} \leqslant \mathcal{S}(\lambda) + (\|Y\|_\infty \vee 1) \log \frac{1}{\delta} \sqrt{\frac{d_\infty(\lambda)}{n}}.$$

Conclusion

Thank you for your attention !