An inverse problem perspective on machine learning

Lorenzo Rosasco

University of Genova
Massachusetts Institute of Technology
Istituto Italiano di Tecnologia
lcsl.mit.edu

Feb 9th, 2018 – Inverse Problems and Machine Learning Workshop, CM+X Caltech

Laboratory for Computational and Statistical Learning

Today selection

- Classics:

  “Learning as an inverse problem”

- Latest releases:

  “Kernel methods as a test bed for algorithm design”

Outline

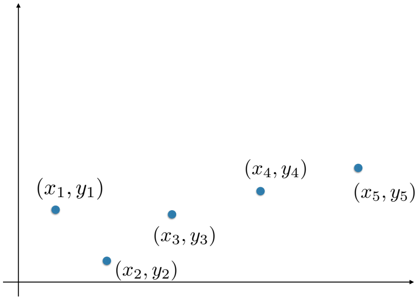
Learning theory 2000

Learning as an inverse problem

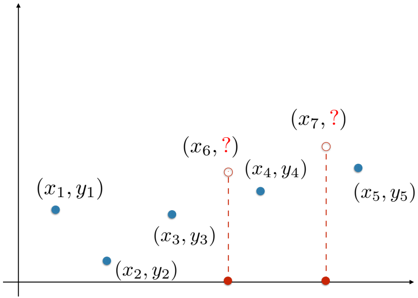
Regularization

Recent advances

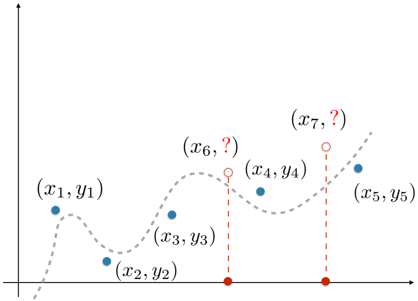
What's learning



What's learning



What's learning



Learning is about inference not interpolation

Statistical Machine Learning (ML)

- (X, Y) a pair of random variables in $\mathcal{X} \times \mathbb{R}$.
- $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ a loss function.
- $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$

Problem: Solve

$$\min_{f \in \mathcal{H}} \mathbb{E}[L(f(X), Y)]$$

given only $(x_1, y_1), \ldots, (x_n, y_n)$, a sample of n i.i. copies of (X, Y).

ML theory around 2000-2010

- All algorithms are ERM (empirical risk minimization)

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i)$$

[Vapnik '96]

- Emphasis on empirical process theory...

$$\mathbb{P}\left(\sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} L(f(X_i), Y_i) - \mathbb{E}[L(f(X), Y)] \right| > \epsilon \right)$$

[Vapnik, Chervonenkis,'71 Dudley, Giné, Zinn '94]

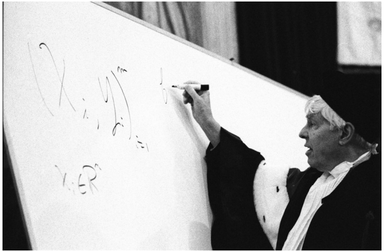
- ...and complexity measures, e.g. Gaussian/Rademacher complexities

$$C(\mathcal{H}) = \mathbb{E} \sup_{f \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i f(X_i)$$

[Barlett, Bousquet, Koltchinskii, Massart, Mendelson... 00]

Around the same time

Cucker and Smale, On the mathematical foundations of learning theory, AMS



- Caponnetto, De Vito and R. Verri, **Learning as an Inverse Problem**, JMLR
- Smale, Zhou, Shannon sampling and function reconstruction from point values, Bull. AMS

Outline

Learning theory 2000

Learning as an inverse problem

Regularization

Recent advances

Inverse Problems (IP)

- $A : \mathcal{H} \to \mathcal{G}$ bounded linear operator, between Hilbert spaces
- $g \in \mathcal{G}$

Problem: Find f solving

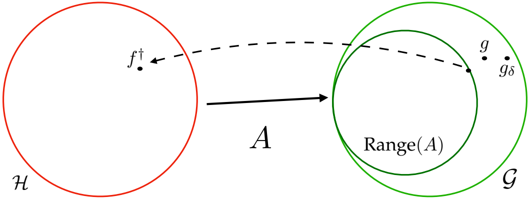
$$Af = g$$

assuming A and g_δ are given, with $\|g - g_\delta\| \leq \delta$

[Engl, Hanke, Neubauer'96]

Ill-posedeness

- Existence: $g \notin \text{Range}(A)$
- Uniqueness: $\text{Ker}(A) \neq \emptyset$
- Stability: $\|A^\dagger\| = \infty$ (large is also a mess)



$$\mathcal{O} = \underset{\mathcal{H}}{\text{argmin}}\,\|Af - g\|^2, \qquad f^\dagger = A^\dagger g = \min_{\mathcal{O}}\|f\|$$

Is machine learning an inverse problem?

- (X, Y)
- $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$
- $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$

Solve

$$\min_{f \in \mathcal{H}} \mathbb{E}[L(f(X), Y)]$$

given only $(x_1, y_1), \ldots, (x_n, y_n)$.

- $A : \mathcal{H} \to \mathcal{G}$
- $g \in \mathcal{G}$

Find f solving

$$Af = g$$

given A and g_δ with $\|g - g_\delta\| \leq \delta$

Actually yes, under some assumptions.

Key assumptions: least squares and RKHS

Assumption

$$L(f(x), y) = (f(x) - y)^2$$

Assumption

- *$(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is a Hilbert space (real, separable)*
- *continuous evaluation functionals, for all $x \in \mathcal{X}$, let $e_x : \mathcal{H} \to \mathbb{R}$, with $e_x(f) = f(x)$, then*

$$|e_x(f) - e_x(f')| \lesssim \|f - f'\|$$

[Aronszajn '50]

Key assumptions: least squares and RKHS

Assumption

$$L(f(x), y) = (f(x) - y)^2$$

Assumption

- *$(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is a Hilbert space (real, separable)*
- *continuous evaluation functionals, for all $x \in \mathcal{X}$, let $e_x : \mathcal{H} \to \mathbb{R}$, with $e_x(f) = f(x)$, then*

$$|e_x(f) - e_x(f')| \lesssim \|f - f'\|$$

Implications

[Aronszajn '50]

- $\|f\|_\infty \lesssim \|f\|$
- $\exists\ k_x \in \mathcal{H}$ such that

$$f(x) = \langle f, k_x \rangle$$

Interpolation and sampling operator

[Bertero, De mol, Pike '85,'88]

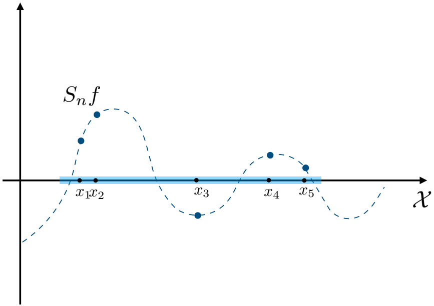
$$f(x_i) = \langle f, k_{x_i} \rangle = y_i, \quad i = 1, \dots, n$$

$$\Downarrow$$

$$S_n f = \mathbf{y}$$

Sampling operator: $S_n : \mathcal{H} \to \mathbb{R}^n$,

$$(S_n f)^i = \langle f, k_{x_i} \rangle \, , \quad \forall i = 1, \dots, n$$



Learning and restriction operator

[Caponnetto, De Vito, R. '05]

$$\langle f, k_x \rangle = f_\rho(x), \quad \rho - a.s.$$

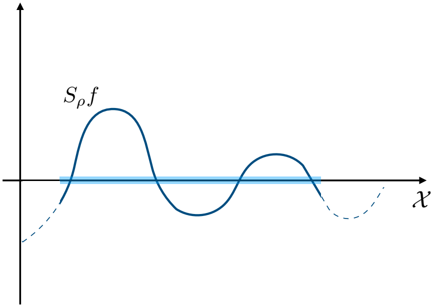
$$\Downarrow$$

$$S_\rho f = f_\rho$$

$f_\rho(x) = \int d\rho(x,y) y$ ρ-almost surely.
$L^2(\mathcal{X}, \rho) = \{ f \in \mathbb{R}^{\mathcal{X}} \mid \|f\|_\rho^2 = \int d\rho |f(x)|^2 < \infty \}$

Restriction operator: $S_\rho : \mathcal{H} \to L^2(\mathcal{X}, \rho)$,

$$(S_\rho f)(x) = \langle f, k_x \rangle, \quad \rho - \text{almost surely.}$$



Learning as an inverse problem

Inverse problem

Find f solving

$$S_\rho f = f_\rho$$

given S_n and $\mathbf{y}_n = (y_1, \ldots, y_n)$.

Learning as an inverse problem

Inverse problem

Find f solving

$$S_\rho f = f_\rho$$

given S_n and $\mathbf{y}_n = (y_1, \dots, y_n)$.

Least squares

$$\min_{\mathcal{H}} \|S_\rho f - f_\rho\|_\rho^2, \qquad \|S_\rho f - f_\rho\|_\rho^2 = \mathbb{E}(f(X) - Y)^2 - \mathbb{E}(f_\rho(X) - Y)^2$$

Let's see what we got

- Noise model
- Integral operators & covariance operators
- Kernels

Noise model

Ideal

$$S_\rho f = f_\rho$$

$$S_\rho^* S_\rho f = S_\rho^* f_\rho$$

Empirical

$$S_n f = \mathbf{y}$$

$$S_n^* S_n f = S_n^* \mathbf{y}$$

Noise model

$$\|S_n^* \mathbf{y} - S_\rho^* f_\rho\| \leq \delta_1 \qquad \|S_\rho^* S_\rho - S_n^* S_n\| \leq \delta_2$$

Inverse problem discretization, Econometrics

Integral and covariance operators operators

- Extension operator $S_\rho^* : L^2(\mathcal{X}, \rho) \to \mathcal{H}$

$$S_\rho^* f(x') = \int d\rho(x) k(x', x) f(x)$$

where $k(x, x') = \langle k_x, k'_x \rangle$ is pos.def.

- Covariance operator $S_\rho^* S_\rho : \mathcal{H} \to \mathcal{H}$

$$S_\rho^* S_\rho = \int d\rho(x) k_x \otimes k_{x'}$$

Kernels

Choosing a RKHS implies choosing a representation.

Theorem (Moore-Aronzaijn)

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*, pos.def., then the completion of*

$$\{f \in \mathbb{R}^{\mathcal{X}} \mid f = \sum_{j=1}^{N} c_i k_{x_i},\ c_1, \ldots, c_N \in \mathbb{R},\ x_1, \ldots, x_N \in \mathcal{X}, N \in \mathbb{N}\}$$

w.r.t. $\langle k_x, k'_x \rangle = k(x, x')$ *is a RKHS.*

Kernels

If $K(x, x') = x^\top x'$, then,

- S_n is the n by D data matrix (S_ρ <u>infinite</u> data matrix)
- $S_n^* S_n$ and $S_\rho^* S_\rho$ are the empirical and true covariance operators

Kernels

If $K(x, x') = x^\top x'$, then,

- S_n is the n by D data matrix (S_ρ infinite data matrix)
- $S_n^* S_n$ and $S_\rho^* S_\rho$ are the empirical and true covariance operators

Other kernels:

- $K(x, x') = (1 + x^\top x')^p$
- $K(x, x') = e^{-\|x-x'\|^2 \gamma}$
- $K(x, x') = e^{-\|x-x'\| \gamma}$

What now?

Steal

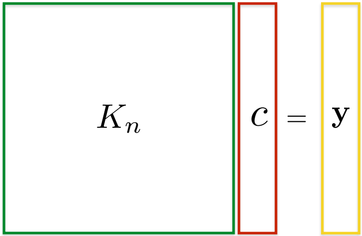
Outline

Tikhonov aka ridge regression

$$f_n^\lambda = (S_n^* S_n + \lambda n I)^{-1} S_n^* \mathbf{y}$$

Tikhonov aka ridge regression

$$f_n^\lambda = (S_n^* S_n + \lambda n I)^{-1} S_n^* \mathbf{y} = S_n^* (\underbrace{S_n S_n^*}_{K_n} + \lambda n I)^{-1} \mathbf{y}$$



Statistics

Theorem (Caponnetto De Vito '05)

Assume $K(X,X), |Y| \leq 1$ *a.s. and* $f^\dagger \in$ *Range*$(S_\rho S_\rho^*)^r$, $1/2 < r < 1$. *If* $\lambda_n = n^{-\frac{1}{2r+1}}$

$$\mathbb{E}[\|Sf_n^{\lambda_n} - f^\dagger\|_\rho^2] \lesssim n^{-\frac{2r}{2r+1}}$$

Statistics

Theorem (Caponnetto De Vito '05)

Assume $K(X,X), |Y| \leq 1$ *a.s. and* $f^\dagger \in$ *Range*$(S_\rho S_\rho^*)^r$, $1/2 < r < 1$. *If* $\lambda_n = n^{-\frac{1}{2r+1}}$

$$\mathbb{E}[\|Sf_n^{\lambda_n} - f^\dagger\|_\rho^2] \lesssim n^{-\frac{2r}{2r+1}}$$

Proof

$$\forall \lambda > 0, \qquad \mathbb{E}[\|Sf_n^{\lambda} - f_\rho\|_\rho^2] \lesssim \frac{1}{\lambda}(\delta_1 + \delta_2) \; + \; \lambda^{2r}$$

$$\mathbb{E}[\delta_1], \mathbb{E}[\delta_2] \lesssim \frac{1}{\sqrt{n}}$$

Iterative regularization

From the Neumann series...

$$f_n^t = \gamma \sum_{j=0}^{t-1} (I - \gamma S_n^* S_n)^j S_n^* \mathbf{y}$$

Iterative regularization

From the Neumann series. . .

$$f_n^t = \gamma \sum_{j=0}^{t-1} (I - \gamma S_n^* S_n)^j S_n^* \mathbf{y} = \gamma S_n^* \sum_{j=0}^{t-1} (I - \gamma \underbrace{S_n S_n^*}_{K_n})^j \mathbf{y}$$

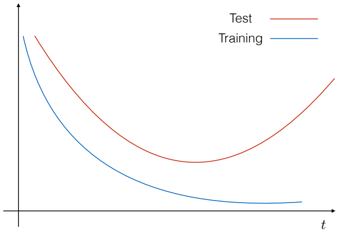
Iterative regularization

From the Neumann series...

$$f_n^t = \gamma \sum_{j=0}^{t-1} (I - \gamma S_n^* S_n)^j S_n^* \mathbf{y} = \gamma S_n^* \sum_{j=0}^{t-1} (I - \gamma \underbrace{S_n S_n^*}_{K_n})^j \mathbf{y}$$

...to gradient descent

$$f_n^t = f_n^{t-1} - \gamma S_n^* (S_n f_n^{t-1} - \mathbf{y}) \qquad c_n^t = c_n^{t-1} - \gamma (K_n c_n^{t-1} - \mathbf{y})$$



Iterative regularization statistics

Theorem (Bauer, Pereverzev, R. '07)

Assume $K(X,X), |Y| \leq 1$ *a.s. and* $f^\dagger \in$ *Range*$(S_\rho S_\rho^*)^r$, $1/2 < r < \infty$. *If* $t_n = n^{\frac{1}{2r+1}}$

$$\mathbb{E}[\|Sf_n^{t_n} - f^\dagger\|_\rho^2] \lesssim n^{-\frac{2r}{2r+1}}$$

Iterative regularization statistics

Theorem (Bauer, Pereverzev, R. '07)

Assume $K(X,X), |Y| \leq 1$ *a.s. and* $f^\dagger \in$ *Range*$(S_\rho S_\rho^*)^r$, $1/2 < r < \infty$. *If* $t_n = n^{\frac{1}{2r+1}}$

$$\mathbb{E}[\|S f_n^{t_n} - f^\dagger\|_\rho^2] \lesssim n^{-\frac{2r}{2r+1}}$$

Proof

$$\forall \lambda > 0, \qquad \mathbb{E}[\|S f_n^t - f_\rho\|_\rho^2] \lesssim \; t\,(\delta_1 + \delta_2) \; + \; \frac{1}{t^{2r}}$$

$$\mathbb{E}[\delta_1], \mathbb{E}[\delta_2] \lesssim \frac{1}{\sqrt{n}}$$

Tikhonov vs iterative regularization

- Same statistical properties...
- ... but time complexities are different $O(n^3)$ vs $O(n^2 n^{\frac{1}{2r+1}})$,
- Iterative regularization provides a bridge between statistics and computations.
- Kernel methods become a test bed for algorithmic solutions.

Computational regularization

Tikhonov

time $O(n^3)$ + space $O(n^2)$ for $1/\sqrt{n}$ learning bound

Computational regularization

Tikhonov

time $O(n^3)$ + space $O(n^2)$ for $1/\sqrt{n}$ learning bound

$$\Downarrow$$

Iterative regularization

time $O(n^2\sqrt{n})$ + space $O(n^2)$ for $1/\sqrt{n}$ learning bound

Outline

Learning theory 2000

Learning as an inverse problem

Regularization

Recent advances

Steal from optimization

Acceleration

- Conjugate gradient

[Blanchard, Kramer '96]

- Chebyshev method

[Bauer, Pervezev. R. '07]

- Nesterov acceleration (Nesterov, '83)

[Salzo, R. '18]

Stochastic gradient

- Single pass stochastic gradient

[Tarres, Yao, '05, Pontil, Ying, '09, Bach, Dieuleveut, Flammarion, '17]

- Multi-pass incremental gradient

[Villa, R. '15]

- Multi-pass stochastic gradient with mini-batches.

[Lin, R. '16]

Computational regularization

Iterative regularization

time $O(n^2\sqrt{n})$ + space $O(n^2)$ for $1/\sqrt{n}$ learning bound

$$\Downarrow$$

Stochastic iterative regularization

time $O(n^2)$ + space $O(n^2)$ for $1/\sqrt{n}$ learning bound

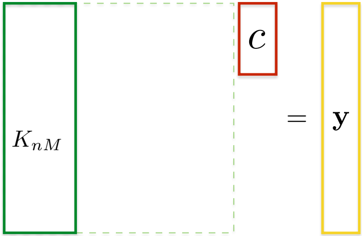
Can we do better? How about memory?

Regularization with projection and preconditioning

[Halko, Martinsson, Tropp '09]

$$(K_{nM}^\top K_{nM} + \lambda n K_{MM})c = K_{nM}^\top \mathbf{y}$$

$$BB^\top = \left(\frac{n}{M} K_{MM}^2 + \lambda n K_{MM}\right)^{-1}$$



FALKON [Rudi, Carratino, R. '17], see also [Ma, Belkin '17]

$$c_t = B\beta_t$$

$$\beta_t = \beta_{t-1} - \frac{\gamma}{n} B^\top \left[K_{nM}^\top (K_{nM} B\beta_{t-1} - \mathbf{y}) \ + \ \lambda n K_{MM} B\beta_{t-1}\right]$$

Falkon statistics

Theorem (Rudi, Carratino, R. '17)

Assume $K(X,X), |Y| \leq 1$ *a.s. and* $f^\dagger \in$ *Range*$(S_\rho S_\rho^*)^r$, $1/2 < r < \infty$. *If*

$$\lambda_n = n^{-\frac{1}{2r+1}}, \qquad M_n = n^{\frac{1}{2r+1}}, \qquad t_n = \log n$$

then

$$\mathbb{E}[\|Sf_n^{\lambda_n, t_n, M_n} - f^\dagger\|_\rho^2] \lesssim n^{-\frac{2r}{2r+1}}$$

Computational regularization

time $O(n^2)$ + space $O(n^2)$ for $1/\sqrt{n}$ learning bound

$$\Downarrow$$

time $\tilde{O}(n\sqrt{n})$ + space $O(n\sqrt{n})$ for $1/\sqrt{n}$ learning bound

Some results

	MillionSongs			YELP		TIMIT	
	MSE	Relative error	Time(s)	RMSE	Time(m)	c-err	Time(h)
FALKON	**80.30**	$\mathbf{4.51 \times 10^{-3}}$	**55**	**0.833**	**20**	32.3%	**1.5**
Prec. KRR	-	4.58×10^{-3}	$289^{\dagger}$	-	-	-	-
Hierarchical	-	4.56×10^{-3}	$293^{\star}$	-	-	-	-
D&C	80.35	-	737^{*}	-	-	-	-
Rand. Feat.	80.93	-	772^{*}	-	-	-	-
Nyström	80.38	-	876^{*}	-	-	-	-
ADMM R. F.	-	5.01×10^{-3}	$958^{\dagger}$	-	-	-	-
BCD R. F.	-	-	-	0.949	$42^{\ddagger}$	34.0%	$1.7^{\ddagger}$
BCD Nyström	-	-	-	0.861	$60^{\ddagger}$	33.7%	$1.7^{\ddagger}$
KRR	-	4.55×10^{-3}	-	0.854	$500^{\ddagger}$	33.5%	$8.3^{\ddagger}$
EigenPro	-	-	-	-	-	32.6%	$3.9^{\wr}$
Deep NN	-	-	-	-	-	32.4%	-
Sparse Kernels	-	-	-	-	-	**30.9%**	-
Ensemble	-	-	-	-	-	33.5%	-

Conclusions

Contribution

- Learning as an inverse problems
- Computational regularization: statistics meets numerics

Future work

- Scaling things up...
- Regularization with projections (quadrature, Galerkin methods)
- Connection to PDE/integral equations: exploit more structure
- Structured prediction/deep learning
- Semisupervised/unsupervised learning
- Embedding and compressed learning