Learning with random features

Alessandro Rudi
INRIA - École Normale Supérieure, Paris

joint work with Lorenzo Rosasco (IIT-MIT)

January 17th, 2018 – Cambridge

Data+computers+ machine learning = AI/Data science

- 1Y US data center= 1M houses
- MobileEye pays 1000 labellers

Can we make do with less?

Beyond a theoretical divide
→ Integrate statistics and numerics/optimization

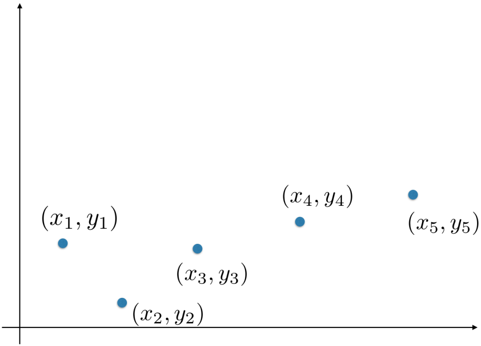
Outline

Part I: Random feature networks

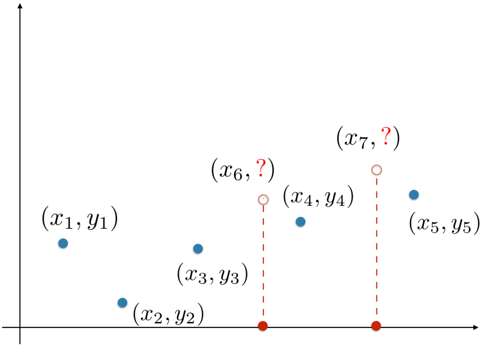
Part II: Properties of RFN

Part III: Refined results on RFN

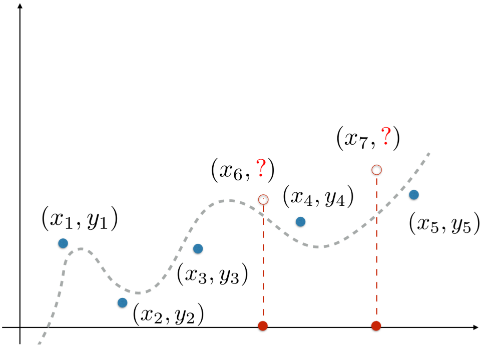
Supervised learning



Supervised learning



Supervised learning



Problem: **given** $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ **find** $f(x_{\text{new}}) \sim y_{\text{new}}$

Neural networks

$$f(x) = \sum_{j=1}^{M} \beta_j \sigma(w_j^\top x + b_j)$$

- $\sigma : \mathbb{R} \to \mathbb{R}$ a non linear *activation* function.
- For $j = 1, \ldots, M$, β_j, w_j, b_j parameters to be determined.

Neural networks

$$f(x) = \sum_{j=1}^{M} \beta_j \sigma(w_j^\top x + b_j)$$

- $\sigma : \mathbb{R} \to \mathbb{R}$ a non linear *activation* function.
- For $j = 1, \ldots, M$, β_j, w_j, b_j parameters to be determined.

Some references

- **History** [McCulloch, Pitts '43; Rosenblatt '58; Minsky, Papert '69; Y. LeCun, '85; Hinton et al. '06]
- **Deep learning** [Krizhevsky et al. '12 - 18705 Cit.!!!]
- **Theory** [Barron '92-94; Bartlett, Anthony '99; Pinkus, '99]

Random features networks

$$f(x) = \sum_{j=1}^{M} \beta_j \sigma({w_j}^\top x + b_j)$$

- $\sigma : \mathbb{R} \to \mathbb{R}$ a non linear *activation* function.
- For $j = 1, \ldots, M$, β_j parameters to be determined
- For $j = 1, \ldots, M$, w_j, b_j chosen at random

Random features networks

$$f(x) = \sum_{j=1}^{M} \beta_j \sigma({w_j}^\top x + b_j)$$

- $\sigma : \mathbb{R} \to \mathbb{R}$ a non linear *activation* function.
- For $j = 1, \dots, M$, β_j parameters to be determined
- For $j = 1, \dots, M$, w_j, b_j chosen at random

Some references

- **Neural nets** [Block '62], **Extreme learning machine** [Huang et al. '06] 5196 Cit.??
- **Sketching/one-bit compressed sensing** see e.g. [Plan, Vershynin '11-14]

$$x \mapsto \sigma(S^\top x), \qquad S \text{ random matrix}$$

- **Gaussian processes/kernel methods** [Neal '95, Rahimi, Recht '06'08'08]

From RFN to PD kernels

$$\frac{1}{M}\sum_{j=1}^{M}\sigma(w_j^\top x+b_j)\sigma(w_j^\top x'+b_j) \approx K(x,x') = \mathbb{E}[\sigma(W^\top x+B)\sigma(W^\top x'+B)]$$

From RFN to PD kernels

$$\frac{1}{M}\sum_{j=1}^{M}\sigma(w_j^\top x+b_j)\sigma(w_j^\top x'+b_j) \approx K(x,x') = \mathbb{E}[\sigma(W^\top x+B)\sigma(W^\top x'+B)]$$

Example I: Gaussian kernel/Random Fourier features [Rahimi, Recht '08]

Let $\sigma(\cdot) = \cos(\cdot)$, $W \sim N(0, I)$ and $B \sim U[0, 2\pi]$

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

Example II: Arccos kernel/ReLU features [Le Roux, Bengio '07; Chou, Saul '09]

Let $\sigma(\cdot) = |\cdot|_+$, $(W, B) \sim U[\mathbb{S}^{d+1}]$

$$K(x,x') = \sin\theta + (\pi-\theta)\cos\theta, \qquad \theta = \mathsf{arcos}(x^\top x')$$

A general view

Let X a measurable space and $K : X \times X \to \mathbb{R}$ symmetric and pos. def.

Assumption (RF)

There exist

- W *random var. in* $\mathcal{W}$ *with law* π*.*
- $\phi : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ *a measurable function.*

such that for all $x, x' \in X$*,*

$$K(x, x') = \mathbb{E}[\phi(W, x)\phi(W, x')].$$

A general view

Let X a measurable space and $K : X \times X \to \mathbb{R}$ symmetric and pos. def.

Assumption (RF)

There exist

- W *random var. in* $\mathcal{W}$ *with law* π.
- $\phi : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ *a measurable function.*

such that for all $x, x' \in X$,

$$K(x, x') = \mathbb{E}[\phi(W, x)\phi(W, x')].$$

Random feature representation Given a sample $w_1, \dots, w_M$ of M i.i. copies of W consider

$$K(x, x') \approx \frac{1}{M} \sum_{j=1}^{M} \phi(w_j, x)\phi(w_j, x')$$

Functional view

Reproducing kernel Hilbert space (RKHS) **[Aronzaijn '50]**: $\mathcal{H}_K$ space of functions

$$f(x) = \sum_{j=1}^{p} \beta_j K(x, x_j)$$

completed with respect to $\langle K_x, K'_x \rangle := K(x, x')$.

Functional view

Reproducing kernel Hilbert space (RKHS) [Aronzaijn '50]: $\mathcal{H}_K$ space of functions

$$f(x) = \sum_{j=1}^{p} \beta_j K(x, x_j)$$

completed with respect to $\langle K_x, K'_x \rangle := K(x, x')$.

RFN spaces: $\mathcal{H}_{\phi,p}$ space of functions

$$f(x) = \int d\pi(w)\beta(w)\phi(w, x),$$

with $\|\beta\|_p^p = \mathbb{E}|\beta(W)|^p < \infty$.

Functional view

Reproducing kernel Hilbert space (RKHS) [Aronzaijn '50]: $\mathcal{H}_K$ space of functions

$$f(x) = \sum_{j=1}^{p} \beta_j K(x, x_j)$$

completed with respect to $\langle K_x, K_x' \rangle := K(x, x')$.

RFN spaces: $\mathcal{H}_{\phi,p}$ space of functions

$$f(x) = \int d\pi(w)\beta(w)\phi(w, x),$$

with $\|\beta\|_p^p = \mathbb{E}|\beta(W)|^p < \infty$.

Theorem (Schoenberg, '38, Aronzaijn '50)

Under Assumption (RF), Then,

$$\mathcal{H}_K \simeq \mathcal{H}_{\phi,2}.$$

Why should you care

RFN promises

- Replace optimization with randomization in NN.
- Reduce memory/time footprint of GP/kernel methods.

Outline

Kernel approximations

$$\begin{aligned}\tilde{K}(x,x') &= \frac{1}{M}\sum_{j=1}^{M}\phi(w_j,x)\phi(w_j,x') \\ K(x,x') &= \mathbb{E}[\phi(W,x)\phi(W,x')]\end{aligned}$$

Theorem

Assume ϕ is bounded. Let $\mathcal{K} \subset X$ compact, then w.h.p.

$$\sup_{x\in\mathcal{K}} |K(x,x) - \tilde{K}(x,x)| \lesssim \frac{C_{\mathcal{K}}}{\sqrt{M}}$$

- [Rahimi, B. Recht '08, Sutherland, Schneider '15 , Sriperumbudur, Szabó '15]
- Empirical characteristic function [Feuerverger, Mureika '77, Csörgó '84, Yukich '87]

Supervised learning

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

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).

Rahimi & Recht estimator

Ideally, $\mathcal{H} = \mathcal{H}_{\phi,\infty,R}$, the space of functions

$$f(x) = \int d\pi(w)\beta(w)\phi(w,x), \qquad \|\beta\|_\infty \leq R.$$

In practice, $\mathcal{H} = \mathcal{H}_{\phi,\infty,R,M}$ the space of functions

$$f(x) = \sum_{j=1}^{M} \tilde{\beta}_j \phi(w_j, x), \qquad \sup_j |\tilde{\beta}_j| \leq R.$$

Estimator

$$\operatorname*{argmin}_{f \in \mathcal{H}_{\phi,\infty,R,M}} \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i)$$

Rahimi & Recht result

Theorem (Rahimi, Recht '08)

Assume L is ℓ-Lipschitz and convex. If ϕ is bounded, then w.h.p.

$$L(\widehat{f}(X), Y)] - \min_{f \in \mathcal{H}_{\phi,\infty,R}} \mathbb{E}[L(f(X), Y)] \lesssim \ell R \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{M}} \right)$$

Other result: [Bach '15], replaced $\mathcal{H}_{\phi,\infty,R}$ with a ball in $\mathcal{H}_{\phi,2}$.

R needs be fixed and $M = n$ is needed for $1/\sqrt{n}$ rates.

Our approach

For $f_\beta(x) = \sum_{j=1}^M \beta_j \phi(w_j, x)$, consider

RF-ridge regression

$$\min_{\beta \in \mathbb{R}^M} \frac{1}{n} \sum_{i=1}^n (y_i - f_\beta(x_i))^2 + \lambda \sum_{j=1}^M |\beta_j|^2$$

Our approach

For $f_\beta(x) = \sum_{j=1}^M \beta_j \phi(w_j, x)$, consider

RF-ridge regression

$$\min_{\beta \in \mathbb{R}^M} \frac{1}{n} \sum_{i=1}^n (y_i - f_\beta(x_i))^2 + \lambda \sum_{j=1}^M |\beta_j|^2$$

Computations

$$\widehat{\beta}_\lambda = (\widehat{\Phi}\widehat{\Phi} + \lambda n I)^{-1} \widehat{\Phi}^\top \widehat{y}$$

- $\widehat{\Phi}_{i,j} = \phi(w_j, x_i)$, $n \times M$ data matrix
- $\widehat{y}$ $n \times 1$ outputs vector

Computational footprint

$$\widehat{\beta}_\lambda = (\widehat{\Phi}^\top \widehat{\Phi} + \lambda n I)^{-1} \widehat{\Phi}^\top \widehat{y}$$

$O(nM^2)$ time and $O(Mn)$ memory cost

Compare to $O(n^3)$ and $O(n^2)$ using kernel methods/GP.

What are the learning properties if $M < n$?

Worst case: basic assumptions

Noise

$$\mathbb{E}[|Y|^p \mid X = x] \leq \frac{1}{2} p! \sigma^2 b^{p-2}, \quad \forall p \geq 2$$

RF boundness: Under assumption (RF), let ϕ be bounded.

Best model: There exists $f^\dagger$ solving

$$\min_{f \in \mathcal{H}_{\phi,2}} \mathbb{E}[(Y - f(X))^2].$$

Note:
- we allow to consider the whole space $\mathcal{H}_{\phi,2}$ rather than a ball.
- We allow misspecified models (regression function $\notin \mathcal{H}$).

Worst case: analysis

Theorem (Rudi, R. '17)

Under the basic assumptions, let $\widehat{f} = f_{\widehat{\beta}_\lambda}$ *then w.h.p.*

$$\mathbb{E}[(Y - \widehat{f}_\lambda(X))^2] - \mathbb{E}[(Y - f^\dagger(X))^2] \lesssim \frac{1}{\lambda n} + \lambda + \frac{1}{M},$$

so that, for

$$\widehat{\lambda} = O\left(\frac{1}{\sqrt{n}}\right), \qquad \widehat{M} = O\left(\frac{1}{\widehat{\lambda}}\right)$$

then w.h.p.

$$\mathbb{E}[(Y - \widehat{f}_{\widehat{\lambda}}(X))^2] - \mathbb{E}[(Y - f^\dagger(X))^2] \lesssim \frac{1}{\sqrt{n}}.$$

Remarks

- Match statistical minmax lower bounds [Caponnetto, De Vito '05].
- Special case: Sobolev spaces with $s = 2d$, e.g. exponential kernel and Fourier features.
- Corollaries for classification using plugin classifiers [Audibert, Tsybakov '07; Yao, Caponnetto, R. '07]
- Same statistical bound of (kernel) ridge regression [Caponnetto, De Vito '05].

$$\boxed{M = \sqrt{n} \text{ suffices for } \tfrac{1}{\sqrt{n}} \text{ rates.}}$$

$O(n^2)$ time and $O(n\sqrt{n})$ memory suffice, rather than $O(n^3)/O(n^2)$

Some ideas from the proof

[Caponnetto, De Vito, R. '05- , Smale, Zhou'05]

Fixed design linear regression

$$\widehat{y} = \widehat{X} w_* + \delta$$

Ridge regression

$$\begin{aligned}
&\widehat{X}(\widehat{X}^\top \widehat{X} + \lambda)^{-1}\widehat{y} - \widehat{X} w_* = \\
&\widehat{X}\widehat{X}^\top(\widehat{X}^\top \widehat{X}\top + \lambda)^{-1}\delta - \widehat{X}((\widehat{X}^\top \widehat{X} + \lambda)^{-1} - I)w_* = \\
&\widehat{X}\widehat{X}^\top(\widehat{X}^\top \widehat{X}\top + \lambda)^{-1}\delta + \lambda\widehat{X}(\widehat{X}^\top \widehat{X} + \lambda)^{-1}w_*
\end{aligned}$$

Key quantities

$$Lf(x) = \mathbb{E}[K(x,X)f(X)], \qquad L_M f(x) = \mathbb{E}[K_M(x,X)f(X)].$$

Let $K_x = K(x,\cdot)$.

- Noise: $(L_M + \lambda I)^{-\frac{1}{2}} \tilde{K}_X Y$

  [Pinelis '94]

- Sampling: $(L_M + \lambda I)^{-\frac{1}{2}} \tilde{K}_X \otimes \tilde{K}_X$

  [Tropp '12, Minsker '17]

- Bias: $\lambda (L + \lambda I)^{-1} L^{\frac{1}{2}}$

  [. . .]

Key quantities (cont.)

RF approximation:

- $L^{1/2}[(L+\lambda I)^{-1}L - (L_M + \lambda I)^{-1}L_M]$

  [Rudi, R. '17]

- $(I-P)\phi(w,\cdot)$, where $P = L^{\dagger}L$

  [Rudi, R. '17, De Vito; R., Toigo '14]

Note: it can be that $\phi(w,\cdot) \notin \mathcal{H}_k$

Key lemma

Lemma (Rudi, R. '17)

W.h.p.

$$\|L^{1/2}[\underbrace{(L+\lambda I)^{-1}L}_{P_\lambda} - \underbrace{(L_M+\lambda I)^{-1}L_M}_{P_{\lambda,M}}]\| \leq \frac{1}{\sqrt{M}}.$$

Perhaps one might have guessed $1/\lambda M$ or $1/\sqrt{\lambda M}$ from

$$\|P_N^A - P_N^B\| \leq \frac{\|(I-P_N^A)(A-B)P_N^B\|}{\mathsf{gap}_N(A)} \leq \frac{\|A-B\|}{\mathsf{gap}_N(A)}$$

Using ideas from [Rudi, Canas, R. '13]

$O(n^2)$ time and $O(n\sqrt{n})$ memory suffice for $\frac{1}{\sqrt{n}}$ rates.

Is it possible to do better? (Less feature? Better rates?)

Outline

Part I: Random feature networks

Part II: Properties of RFN

Part III: Refined results on RFN

Regularity conditions I: Capacity

Let

$$\mathcal{N}(\lambda) = \mathsf{Trace}((L + \lambda I)^{-1} L)$$

Assumption (C)

Assume

$$\mathcal{N}(\lambda) = O(\lambda^{-\gamma}), \qquad \gamma \in [0, 1]$$

Some remarks:

- Implied by eigenvalue condition $\sigma_i(L) = O(i^{-\frac{1}{\gamma}})$.
- Equivalent to entropy conditions, for Sobolev kernels $\gamma = d/2s$.
- Other regimes can be considered- e.g. analytic/finite rank kernels.

Regularity conditions II: Sparsity

Let where $f_*(x) = \mathbb{E}[Y|X = x]$.

Assumption (S)

$$f_* \in \mathit{Range}(L^r), \qquad r \geq 1/2$$

Equivalently, let (σ_i, ψ_i) be the eigenvalues and eigenfunction of L,

$$\sum_{j=1}^{\infty} \frac{|\langle f_*, \psi_j \rangle|^2}{\sigma_i^{2r}} < \infty$$

Note: For $r = 1/2$ it is equivalent to existence of $f^\dagger$ [Mercer 1909]

Fast rates for RF-ridge regression

Theorem (Rudi, R. '17)

Under the basic assumptions +(C,S), let $\widehat{f}_\lambda = f_{\widehat{\beta}_\lambda}$ *then w.h.p.*

$$\mathbb{E}[(Y - \widehat{f}_\lambda(X))^2] - \mathbb{E}[(Y - f^\dagger(X))^2] \lesssim \frac{\mathcal{N}(\lambda)}{n} + \lambda^{2r} + \frac{\mathcal{N}(\lambda)^{2r-1}}{\lambda^{2r-1}M},$$

so that, for

$$\widehat{\lambda} = O\left(n^{\frac{1}{2r+\gamma}}\right), \qquad \widehat{M} \;=\; O\left(n^{\frac{1+\gamma(2r-1)}{2r+\gamma}}\right)$$

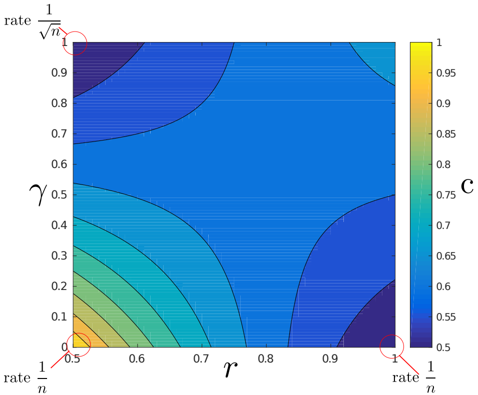
then w.h.p.

$$\mathbb{E}[(Y - \widehat{f}_{\widehat{\lambda}}(X))^2] - \mathbb{E}[(Y - f^\dagger(X))^2] \lesssim \; n^{-\frac{2r}{2r+\gamma}}$$

Remarks

- The obtained rate is minmax optimal [Caponnetto, De Vito '05].
- Reduces to worst case for $\gamma = 1, r = 1/2$.
- $M = O(n)$ in parametric case.

$$M = n^c$$



Adaptive sampling

Leverage scores

- Graph sparsification [Spielman, Srivastava '08]
- Nonparametric regression [Bach '13; Alaoui, Mahoney '15, Rudi, R. '15]

Leverage score RF [Bach '16]

- $s(w) = \mathbb{E}[\phi(X,w)(L+\lambda)^{-1}\phi(X,w)]$
- $C_s := \mathbb{E}\left[s(W)\right]$

Consider

$$\psi_s(x,w) = \psi(x,w)/\sqrt{C_s s(w)},$$

with distribution $\pi_s(w) := \pi(w) C_s s(w)$.

Fast rates for adaptive RF-ridge regression

Theorem (Rudi, R. '17)

Under the basic assumptions+(C,S), let $\widehat{f}_\lambda = f_{\widehat{\beta}_\lambda}$ *then w.h.p.*

$$\mathbb{E}[(Y - \widehat{f}_\lambda(X))^2] - \mathbb{E}[(Y - f^\dagger(X))^2] \lesssim \frac{\mathcal{N}(\lambda)}{n} + \lambda^{2r} + \frac{\lambda \mathcal{N}(\lambda)}{M},$$

so that, for

$$\widehat{\lambda} = O\left(n^{-\frac{1}{2r+\gamma}}\right), \qquad \widehat{M} \;=\; O\left(n^{\frac{\gamma+(2r-1)}{2r+\gamma}}\right)$$

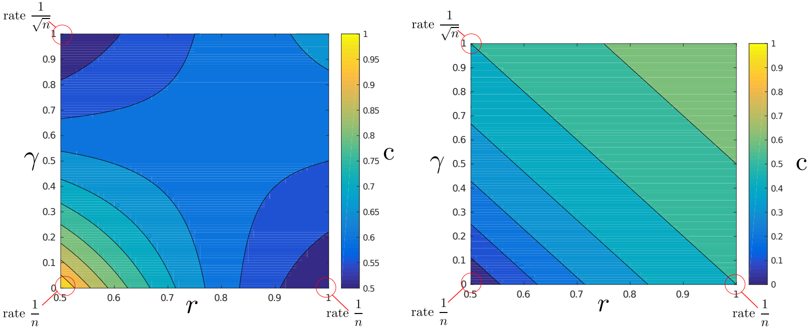
then w.h.p.

$$\mathbb{E}[(Y - \widehat{f}_{\widehat{\lambda}}(X))^2] - \mathbb{E}[(Y - f^\dagger(X))^2] \lesssim \; n^{-\frac{2r}{2r+\gamma}}$$

Remarks

- Same rate as usual
- Much fewer random features! (Compare to [Bach '16])
- $M = O(1)$ in parametric case.

$$M = n^c$$



Contribution

First RF result showing:

computational benefits with no loss of statistical accuracy.

- Add optimization/numerical analysis, see Alessandro's talk on friday.

- (Fast) leverage scores computations
- Othe problems: density estimation,MMD, spectral clustering, kernel-(PCA, ICA, K-means)...,
- Beyond random features: projection methods/Galerkin methods?