Falkon: optimal and efficient large scale kernel learning

Alessandro Rudi
INRIA - École Normale Supérieure

joint work with Luigi Carratino (UniGe), Lorenzo Rosasco (MIT - IIT)

July, 6th – ISMP 2018

Learning problem

The problem $\mathcal{P}$

Find

$$f_{\mathcal{H}} = \operatorname*{argmin}_{f \in \mathcal{H}} \mathcal{E}(f), \qquad \mathcal{E}(f) = \int d\rho(x,y)(y - f(x))^2$$

with ρ **unknown** but given $(x_i, y_i)_{i=1}^n$ i.i.d. samples.

Basic assumtions:

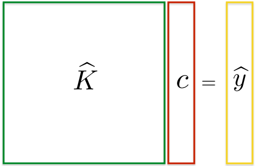
- Tail assumption: $\int |y|^p d\rho \leq \frac{1}{2} p! \sigma^2 b^{p-2}, \quad \forall p \geq 2$
- $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ RKHS with bounded kernel K

Kernel ridge regression

$$\widehat{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2$$

$$\widehat{f}_\lambda(x) = \sum_{i=1}^{n} K(x, x_i) c_i$$

$$(\widehat{K} + \lambda n I)c = \widehat{y}$$



Complexity: **Space** $O(n^2)$ **Kernel eval.** $O(n^2)$ **Time** $O(n^3)$

Random projections

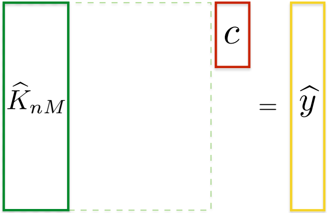
Solve $\widehat{\mathcal{P}}_n$ on $\mathcal{H}_M = \operatorname{span}\{K(\tilde{x}_1, \cdot), \ldots, K(\tilde{x}_M, \cdot)\}$

$$\widehat{f}_{\lambda, M} = \underset{f \in \mathcal{H}_M}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2$$

- ... that is, pick M columns at random

$$\widehat{f}_{\lambda, M}(x) = \sum_{i=1}^{M} K(x, \tilde{x}_i) c_i$$

$$(\widehat{K}_{nM}^{\top} \widehat{K}_{nM} + \lambda n \widehat{K}_{MM}) c = \widehat{K}_{nM}^{\top} \widehat{y}$$



- **Nyström methods** (Smola, Scholköpf '00)
- Gaussian processes: inducing inputs (Quionero-Candela et al '05)
- Galerkin methods and Randomized linear algebra (Halko et al. '11)

Nyström KRR: Statistics (refined)

Let $Lf(x') = \mathbb{E}K(x', x)f(x)$ and $\mathcal{N}(\lambda) = \text{Trace}((L + \lambda I)^{-1}L)$
Capacity condition:

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

Source condition:

$$f_{\mathcal{H}} \in \text{Range}(L^r), \qquad r \geq 1/2$$

Theorem[Rudi, Camoriano, R. '15] Under (basic) and (refined)

$$\mathbb{E}\mathcal{E}(\widehat{f}_{\lambda,M}) - \mathcal{E}(f_{\mathcal{H}}) \lesssim \frac{\mathcal{N}(\lambda)}{n} + \lambda^{2r} + \frac{1}{M}.$$

By selecting $\lambda_n = n^{-\frac{1}{2r+\gamma}}, M_n = \frac{1}{\lambda_n}$

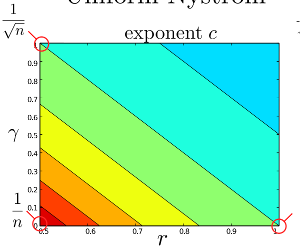
$$\mathbb{E}\mathcal{E}(\widehat{f}_{\lambda_n,M_n}) - \mathcal{E}(f_{\mathcal{H}}) \lesssim n^{-\frac{2r}{2r+\gamma}}$$

Remarks

$$\boxed{M = O(\sqrt{n}) \text{ suffices for } O(1/\sqrt{n}) \text{ rates}}$$

$M = n^c$

Uniform Nyström



- Previous works: only for fixed design
  (Bach ’13, Alaoui, Mahoney, ’15, Yang et al. ’15, Musco, Musco ’16)
- Same minmax bound of KRR [Caponnetto, De Vito ’05].
- Projection regularizes!

Computations required for $O(1/\sqrt{n})$ rate

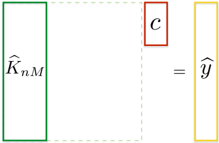
Space:	$O(n)$
Kernel eval.:	$O(n\sqrt{n})$
Time:	$O(n^2)$
Test:	$O(\sqrt{n})$

Possible improvements:

- adaptive sampling
- **optimization**

Optimization to rescue

$$\underbrace{\widehat{K}_{nM}^{\top}\widehat{K}_{nM} + \lambda n \widehat{K}_{MM}}_{H}\, c = \underbrace{\widehat{K}_{nM}^{\top}\widehat{y}}_{b}.$$



Idea: First order methods

$$c_t = c_{t-1} - \frac{\tau}{n}\left[\widehat{K}_{nM}^{\top}(\widehat{K}_{nM}c_{t-1} - y_n) \;+\; \lambda n \widehat{K}_{MM}c_{t-1}\right]$$

Pros: requires $O(nMt)$

Cons: $t \propto \kappa(H)$ arbitrarily large- $\kappa(H) = \sigma_{\max}(H)/\sigma_{\min}(H)$ condition number.

Preconditioning

Idea: solve an equivalent linear system with better condition number

Preconditioning

$$Hc = b \quad \mapsto \quad P^\top H P \beta = P^\top b, \quad c = P\beta.$$

Ideally $PP^\top = H^{-1}$, so that

$$t = O(\kappa(H)) \quad \mapsto \quad t = O(1)!$$

Note: Preconditioning KRR (Fasshauer et al '12, Avron et al '16, Cutajat '16, Ma, Belkin '17)

$$H = K + \lambda n I$$

Can we precondition Nystrom-KRR?

Preconditioning Nystom-KRR

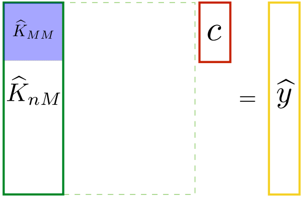
Consider $$H := \widehat{K}_{nM}^{\top}\widehat{K}_{nM} + \lambda n \widehat{K}_{MM}$$

Proposed Preconditioning

$$PP^{\top} = \left(\frac{n}{M}\widehat{K}_{MM}^{2} + \lambda n \widehat{K}_{MM}\right)^{-1}$$

Compare to naive preconditioning

$$PP^{\top} = \left(\widehat{K}_{nM}^{\top}\widehat{K}_{nM} + \lambda n \widehat{K}_{MM}\right)^{-1}.$$



Baby FALKON

Proposed Preconditioning

$$PP^\top = \left(\frac{n}{M}\widehat{K}_{MM}^2 + \lambda n \widehat{K}_{MM}\right)^{-1},$$

Gradient descent

$$\widehat{f}_{\lambda,M,t}(x) = \sum_{i=1}^{M} K(x, \widetilde{x}_i) c_{t,i}, \qquad c_t = P\beta_t$$

$$\beta_t = \beta_{t-1} - \frac{\tau}{n} P^\top \left[\widehat{K}_{nM}^\top (\widehat{K}_{nM} P \beta_{t-1} - y_n) \; + \; \lambda n \widehat{K}_{MM} P \beta_{t-1}\right]$$

FALKON

- Gradient descent $\mapsto$ conjugate gradient
- Computing P

$$P = \frac{1}{\sqrt{n}} T^{-1} A^{-1}, \quad T = \mathrm{chol}(K_{MM}), \quad A = \mathrm{chol}\left(\frac{1}{M}\, TT^\top + \lambda I\right),$$

where $\mathrm{chol}(\cdot)$ is the Cholesky decomposition.

Falkon statistics

Theorem

Under (basic) and (refined), when $M > \frac{\log n}{\lambda}$,

$$\mathbb{E}\mathcal{E}(\widehat{f}_{\lambda_n, M_n, t_n}) - \mathcal{E}(f_{\mathcal{H}}) \lesssim \frac{\mathcal{N}(\lambda)}{n} + \lambda^{2r} + \frac{1}{M} + \exp\left[-t \ \left(1 - \frac{\log n}{\lambda M}\right)^{1/2}\right]$$

By selecting

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

then

$$\mathbb{E}\mathcal{E}(\widehat{f}_{\lambda_n, M_n, t_n}) - \mathcal{E}(f_{\mathcal{H}}) \lesssim n^{-\frac{2r}{2r+\gamma}}$$

Remarks

- Same rates and memory of NKRR, much smaller time complexity, for $O(1/\sqrt{n})$:

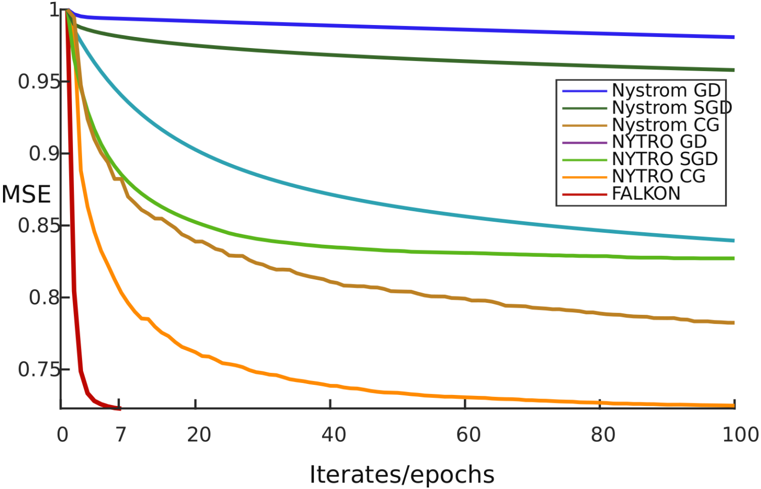
Model:	$O(\sqrt{n})$
Space:	$O(n)$
Kernel eval.:	$O(n\sqrt{n})$
Time:	$\cancel{O(n^2)} \to O(n\sqrt{n})$

Related (worse complexity)

- EigenPro (Belkin et al. '16)
- SGD (Smale, Yao '05, Tarres, Yao '07, Ying, Pontil '08, Bach et al. '14-. . . ,)
- RF-KRR (Rahimi, Recht '07; Bach '15; Rudi, Rosasco '17)
- Divide and conquer (Zhang et al. '13)
- NYTRO (Angles et al '16)
- Nyström SGD (Lin, Rosasco '16)

In practice

Higgs dataset: $n = 10,000,000, \quad M = 50,000$



Some experiments

	MillionSongs ($n \sim 10^6$)			YELP ($n \sim 10^6$)		TIMIT ($n \sim 10^6$)	
	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%	-

Table: MillionSongs, YELP and TIMIT Datasets. Times obtained on: $\ddagger$ = cluster of 128 EC2 r3.2xlarge machines, $\dagger$ = cluster of 8 EC2 r3.8xlarge machines, $\wr$ = single machine with two Intel Xeon E5-2620, one Nvidia GTX Titan X GPU and 128GB of RAM, $\star$ = cluster with 512 GB of RAM and IBM POWER8 12-core processor, $*$ = unknown platform.

Some more experiments

	SUSY ($n \sim 10^6$)			HIGGS ($n \sim 10^7$)		IMAGENET ($n \sim 10^6$)	
	c-err	AUC	Time(*m*)	AUC	Time(*h*)	c-err	Time(*h*)
FALKON	**19.6%**	0.877	**4**	0.833	**3**	20.7%	**4**
EigenPro	19.8%	-	$6^{\wr}$	-	-	-	-
Hierarchical	20.1%	-	$40^{\dagger}$	-	-	-	-
Boosted Decision Tree	-	0.863	-	0.810	-	-	-
Neural Network	-	0.875	-	0.816	-	-	-
Deep Neural Network	-	**0.879**	$4680^{\ddagger}$	**0.885**	$78^{\ddagger}$	-	-
Inception-V4	-	-	-	-	-	**20.0%**	-

Table: Architectures: † cluster with IBM POWER8 12-core cpu, 512 GB RAM, ≀ single machine with two Intel Xeon E5-2620, one Nvidia GTX Titan X GPU, 128GB RAM, ‡ single machine.

Contributions

- Best computations so far for optimal statistics

$$\boxed{\textbf{Space } O(n) \quad \textbf{Time } O(n\sqrt{n})}$$

- In the pipeline: adaptive sampling, general projection, SGD
- TBD other loss, other regularizers, other problems, other solvers...

Proof: bridging statistics and optimization

Lemma

Let $\delta > 0$, $\kappa_P := \kappa(P^\top HP)$, $c_\delta = c_0 \log \frac{1}{\delta}$. *When* $\lambda \geq \frac{1}{n}$

$$\mathcal{E}(\widehat{f}_{\lambda,M,t}) - \mathcal{E}(f_{\mathcal{H}}) \quad \leq \quad \mathcal{E}(\widehat{f}_{\lambda,M}) - \mathcal{E}(f_{\mathcal{H}}) \quad + \quad c_\delta \exp(-t/\sqrt{\kappa_P}).$$

with probability $1 - \delta$.

Lemma

Let $\delta \in (0, 1], \lambda > 0$. *When*

$$M = \frac{2 \log \frac{1}{\delta}}{\lambda},$$

then

$$\kappa(P^\top HP) \leq \left(1 - \frac{\log \frac{1}{\delta}}{\lambda M}\right)^{-1} < 4$$

with probability $1 - \delta$.

Proving $\kappa(P^\top HP) \approx 1$

Let $K_x = K(x, \cdot) \in \mathcal{H}$,

$$C = \int K_x \otimes K_x d\rho_X(x), \quad \widehat{C}_n = \frac{1}{n}\sum_{i=1}^{n} K_{x_i} \otimes K_{x_i}, \quad \widehat{C}_M = \frac{1}{M}\sum_{j=1}^{M} K_{\tilde{x}_j} \otimes K_{\tilde{x}_j}.$$

Recall that $P = \frac{1}{\sqrt{n}} T^{-1}A^{-1}, T = \mathrm{chol}(K_{MM}), A = \mathrm{chol}\left(\frac{1}{M}\ TT^\top + \lambda I\right)$.

Steps

$$1. \quad P^\top HP = A^{-\top}V^*(\widehat{C}_n + \lambda I)VA^{-1}$$

$$2. \quad P^\top HP \ = \ A^{-\top}V^*(\widehat{C}_M + \lambda I)VA^{-1} \ + \ A^{-\top}V^*(\widehat{C}_n - \widehat{C}_M)VA^{-1}$$

$$3. \quad P^\top HP \ = \ I \ + \ A^{-\top}V^*(\widehat{C}_n - \widehat{C}_M)VA^{-1}$$

$$3. \quad P^\top HP \ = \ I \ + \ E \quad \text{with } E = A^{-\top}V^*(\widehat{C}_n - \widehat{C}_M)VA^{-1}$$