FALKON : An Optimal Large Scale Kernel Method

Alessandro Rudi[1], Luigi Carratino[2] and Lorenzo Rosasco[2,3]

[1] INRIA – Sierra-project team & École Normale Supérieure, Paris, France. This work was done when A.R. was working at [3]
[2] DIBRIS – Università degli Studi di Genova, Genoa, Italy
[3] Laboratory for Computational and Statistical Learning – Istituto Italiano di Tecnologia, Genoa, Italy & Massachusetts Institute of Technology, Cambridge, USA

Neural Information Processing Systems Foundation

Motivation

Kernel methods provide a principled way to perform learning with optimal generalization properties. However their high computational requirements limit their applicability to large scale datasets.
FALKON is a novel algorithm that allows to scale up kernel methods preserving optimal statistical accuracy.

LEARNING PROBLEM

Find

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

given $(x_i, y_i)_{i=1}^n$ **i.i.d. samples** from ρ, with ρ **unknown** probability distribution, and $\mathcal{H}$ space of models.

Kernel Ridge Regression

Regularized empirical risk minimization

$n \times n$ matrix s.t. $(\widehat{K})_{ij} = K(x_i, x_j)$

$$\widehat{f}_\lambda = \sum_{i=1}^n \alpha_i K(x, x_i) \qquad (\widehat{K}_{nn} + \lambda n I)\alpha = \widehat{y}$$

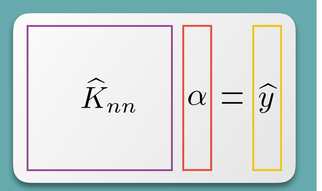
$\widehat{y} = (y_1, \dots, y_n)$

with $K : X \times X \to \mathbb{R}$ positive definite kernel (e.g. $K(x,x') = e^{-\|x-x'\|^2\gamma}$)

COMPLEXITY: $O(n^2)$ space, $O(n^3)$ time

STATISTICS[1]: Let $\lambda_n = 1/\sqrt{n}$, then

$$\mathbb{E}\mathcal{E}(\widehat{f}_{\lambda_n}) - \mathcal{E}(f_{\mathcal{H}}) \lesssim \frac{1}{\sqrt{n}}$$



Nyström Method

Solves regularized ERM on models with **reduced** complexity

$M \ll n$

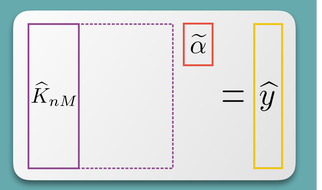
$$\widehat{f}_{\lambda,M} = \sum_{i=1}^M \widetilde{\alpha}_i K(x, \widetilde{x}_i) \quad (\widehat{K}_{nM}^\top \widehat{K}_{nM} + \lambda n \widehat{K}_{MM})\widetilde{\alpha} = \widehat{K}_{MM}^\top \widehat{y} \qquad (1)$$

$\{\widetilde{x}_1, \dots, \widetilde{x}_M\} \subseteq \{x_1, \dots, x_n\}$

COMPLEXITY: $O(M^2)$ space, $O(nM^2)$ time

STATISTICS[2]: Let $\lambda_n = 1/\sqrt{n}$, $M_n = \sqrt{n}$, then

$$\mathbb{E}\mathcal{E}(\widehat{f}_{\lambda_n,M_n}) - \mathcal{E}(f_{\mathcal{H}}) \lesssim \frac{1}{\sqrt{n}}$$



FALKON

Combines **Nyström**, **iterative** optimization and **preconditioning**.

Solves (1) with preconditioner

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

Baby FALKON

$$\widehat{f}_{\lambda,M,t}(x) = \sum_{i=1}^M \alpha_{t,i} K(x, \widetilde{x}_i), \quad \text{with} \quad \alpha_t = B\beta_t$$

$$\beta_k = \beta_{k-1} - \frac{\tau}{n} B^\top \left[\widehat{K}_{nM}^\top (\widehat{K}_{nM}(B\beta_{k-1}) - \widehat{y}) \; + \; \lambda n \widehat{K}_{MM}(B\beta_{k-1})\right]$$

for $t \in \mathbb{N}, \beta_0 = 0$ and $1 \le k \le t$ and a suitable chosen τ.

COMPLEXITY: $O(M^2)$ space, $O(nMt)$ time

The **FINAL ALGORITHM** uses **Conjugate gradient** on preconditioned (1), and computes B as

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

Cholesky decomposition

THEORETICAL ANALISYS

Theorem: Let $\lambda_n = 1/\sqrt{n}, M_n = \sqrt{n}, t = \log(n)$, then

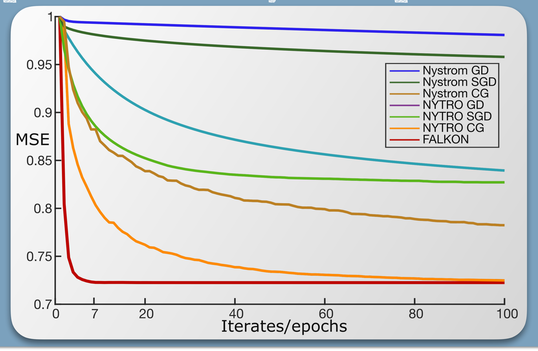
$$\mathbb{E}\mathcal{E}(\widehat{f}_{\lambda_n,M_n,t_n}) - \mathcal{E}(f_{\mathcal{H}}) \lesssim \frac{1}{\sqrt{n}}$$

Optimal generalization is achieved with complexity

$O(n^2) \longrightarrow O(n)$ space,

$O(n^3) \longrightarrow O(n\sqrt{n}\log(n))$ time

Speed/Accuracy Comparison



Test MSE per iterate/epoch over the HIGGS dataset. In **7 iterations** **FALKON** achieves the optimum error value.

Comparison with other methods

Algorithm	train time	kernel evaluations	memory	test time
SVM / KRR + direct method	n^3	n^2	n^2	n
KRR + iterative	$n^2\sqrt[4]{n}$	n^2	n^2	n
Doubly stochastic	$n^2\sqrt{n}$	$n^2\sqrt{n}$	n	n
Pegasos / KRR + sgd	n^2	n^2	n	n
KRR + iter + precond	n^2	n^2	n	n
Divide & Conquer	n^2	$n\sqrt{n}$	n	n
Nyström, random features	n^2	$n\sqrt{n}$	n	$\sqrt{n}$
Nyström + iterative	n^2	$n\sqrt{n}$	n	$\sqrt{n}$
Nyström + sgd	n^2	$n\sqrt{n}$	n	$\sqrt{n}$
FALKON	$\boldsymbol{n\sqrt{n}}$	$\boldsymbol{n\sqrt{n}}$	$\boldsymbol{n}$	$\boldsymbol{\sqrt{n}}$

FALKON achieves the **best time/space complexity** for optimal generalization.

Experiments

	MillionSongs ($n \sim 5\times10^5$)			YELP ($n \sim 10^6$)		TIMIT ($n \sim 10^6$)	
	MSE	Relative error	Time(s)	RMSE	Time(m)	c-err	Time(h)
FALKON	**80.10**	4.51×10^{-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%	$1.7^\ddagger$
EigenPro	-	-	-	-	-	32.6%	$3.9^\wr$
KRR	-	4.55×10^{-3}	-	0.854	$500^\ddagger$	33.5%	$8.3^\ddagger$
Deep NN	-	-	-	-	-	32.4%	-
Sparse Kernels	-	-	-	-	-	**30.9%**	-
Ensemble	-	-	-	-	33.5%	-	

Architectures: ‡ cluster 128 EC2 r3.2xlarge machines; † cluster 8 EC2 r3.8xlarge machines; ≀ single machine with two Intel Xeon E5-2620, one Nvidia GTX Titan X GPU, 128GB RAM; ⋆ cluster with IBM POWER8 12-core processor, 512 GB RAM; ∗ unknown platform.

	SUSY ($n \sim 5\times10^6$)			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%**	-

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

FALKON is coded in MATLAB (on GitHub : "FALKON_paper"). The experiments are run on a **single machine** (two Intel Xeon E5-2630 v3, one NVIDIA Tesla K40c and 128 GB of RAM), while most of the competitors uses clusters of multiple machines.

References

[1] A. Caponnetto and E. De Vito. **Optimal rates for the regularized least-square algorithm**. Foundations of Computational Mathematics 7.3, 2007.
[2] A. Rudi, R. Camoriano and L. Rosasco. **Less is more: Nyström computational regularization.** Advances in Neural Information Processing Systems, 2015.
[3] R. Camoriano, T Angles, A. Rudi and L Rosasco. **NYTRO: When subsampling meets early stopping.** Artificial Intelligence and Statistics, 2016.