(Un)conventional regularization for efficient machine learning

Lorenzo Rosasco
Laboratory for Computational Statistical Learning (LCSL)
University of Genova
Massachusetts Institute of Technology
Istituto Italiano di Tecnologia
lcsl.mit.edu

Nov. 23rd, 2018 – Horizon Maths 2018: Intelligence Artificielle

joint work with L. Carratino, R. Camoriano (LCSL), J. Lin (EPFL), A. Rudi (INRIA)

Laboratory for Computational and Statistical Learning

A general motivation: efficient ML

Beyond the statistics vs optimization dichotomy:

Numerical resources budgeted to data **quality** (not just size)

[Bottou, Bousquet '08]

Outline

Classical regularization

Regularization by optimization

Regularization & Projections

Preconditioning

Statistical learning

Find

$$w_* = \operatorname*{argmin}_{w \in \mathbb{R}^p} \mathbb{E}[(y - \langle w, \Phi(x) \rangle)^2]$$

given only i.i.d. samples $(x_i, y_i)_{i=1}^n$.

We assume throughout that:

- Φ a given high/infinite dimensional representation.
- If $p = \infty$, $\langle \Phi(x), \Phi(x') \rangle$ can be computed in $O(1)$ (kernel methods/Gaussian processes).

Empirical risk minimization (ERM)

$$\widehat{w}_\lambda = \operatorname*{argmin}_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, \Phi(x_i) \rangle)^2 + \lambda \|w\|^2$$

Empirical risk minimization (ERM)

$$\widehat{w}_\lambda = \underset{w \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, \Phi(x_i) \rangle)^2 + \lambda \|w\|^2$$

Theorem (Smale, Zhou '05, Caponetto De Vito '05)

If $p = \infty$, $\|\Phi(x)\|, |y| \leq 1$ *a.s. and* $\lambda = 1/\sqrt{n}$

$$\mathbb{E}[(\langle \Phi(x), \widehat{w}_\lambda \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \frac{1}{\sqrt{n}}$$

Empirical risk minimization (ERM)

$$\widehat{w}_\lambda = \operatorname*{argmin}_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, \Phi(x_i) \rangle)^2 + \lambda \|w\|^2$$

Theorem (Smale, Zhou '05, Caponetto De Vito '05)

If $p = \infty$, $\|\Phi(x)\|, |y| \leq 1$ *a.s. and* $\lambda = 1/\sqrt{n}$

$$\mathbb{E}[(\langle \Phi(x), \widehat{w}_\lambda \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \frac{1}{\sqrt{n}}$$

Proof

$$\forall \lambda > 0, \qquad \mathbb{E}[(\langle \Phi(x), \widehat{w}_\lambda \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \frac{1}{\lambda n} + \lambda$$

Remark: Optimal bound (can be improved under further assumptions)

ERM computations

$$\widehat{w}_\lambda = \operatorname*{argmin}_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, \Phi(x_i) \rangle)^2 + \lambda \|w\|^2$$

Nonparametrics $n < p = \infty$

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

- $\widehat{\Phi}$ is the n by p data/feature matrix
- $\widehat{w}_\lambda = \widehat{\Phi}^\top \widehat{c}_\lambda$
- $\langle \widehat{w}_\lambda, \Phi(x) \rangle = \sum_{i=1}^{n} \langle \Phi(x_i), \Phi(x) \rangle \widehat{c}_\lambda^i$

$$\underbrace{O(n^3)}_{time} + \underbrace{O(n^2)}_{space}$$

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

The rate $1/\sqrt{n}$ is optimal, can we improve time/space requirements?

Outline

Gradient descent (GD) learning

aka L2 boosting, Landweber iteration

GD for

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle w, \Phi(x_i)\rangle)^2 = \frac{1}{n}\|\widehat{\Phi}w - \widehat{y}\|^2$$

Gradient descent (GD) learning

aka L2 boosting, Landweber iteration

GD for

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle w, \Phi(x_i)\rangle)^2 = \frac{1}{n}\|\widehat{\Phi}w - \widehat{y}\|^2$$

Nonparametrics $n \leq p = \infty$

$$\widehat{c}_{t+1} = \widehat{c}_t - \frac{\gamma}{n}(\widehat{\Phi}\widehat{\Phi}^\top \widehat{c}_t - \widehat{y})$$

- $\widehat{w}_\lambda = \widehat{\Phi}^\top \widehat{c}_t$
- $\langle \widehat{w}_t, \Phi(x)\rangle = \sum_{i=1}^{n} \langle \Phi(x_i), \Phi(x)\rangle \, \widehat{c}_t^i$

$$\underbrace{O(n^2 t)}_{time} + \underbrace{O(n^2)}_{space}$$

Gradient descent (GD) learning

aka L2 boosting, Landweber iteration

GD for

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle w, \Phi(x_i)\rangle)^2 = \frac{1}{n}\|\widehat{\Phi}w - \widehat{y}\|^2$$

Nonparametrics $n \leq p = \infty$

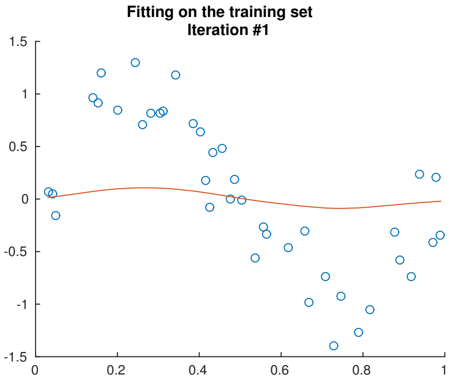
$$\widehat{c}_{t+1} = \widehat{c}_t - \frac{\gamma}{n}(\widehat{\Phi}\widehat{\Phi}^\top \widehat{c}_t - \widehat{y})$$

- $\widehat{w}_\lambda = \widehat{\Phi}^\top \widehat{c}_t$
- $\langle \widehat{w}_t, \Phi(x)\rangle = \sum_{i=1}^{n} \langle \Phi(x_i), \Phi(x)\rangle \, \widehat{c}_t^i$

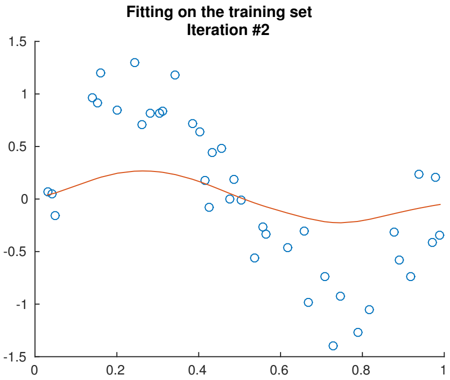
$$\underbrace{O(n^2 t)}_{time} + \underbrace{O(n^2)}_{space}$$

Why should this work??

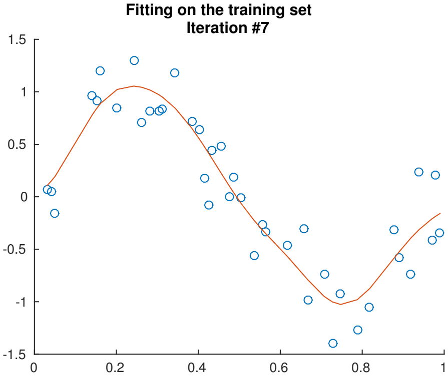
An intuition



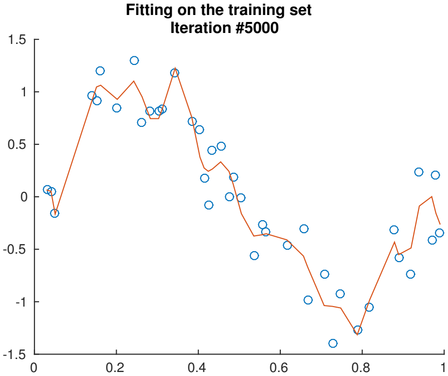
An intuition



An intuition



An intuition



Regularization and stability

$$\widehat{w}_t = \frac{\gamma}{n}\sum_{j=0}^{t-1}(I - \frac{\gamma}{n}\widehat{\Phi}^\top\widehat{\Phi})^j\Phi^\top\widehat{y}$$

Large t

$$\widehat{w}_t = \frac{\gamma}{n}\sum_{j=0}^{t-1}(I - \frac{\gamma}{n}\widehat{\Phi}^\top\widehat{\Phi})^j\Phi^\top\widehat{y} = \quad \approx \quad \frac{\gamma}{n}\sum_{j=0}^{\infty}(I - \frac{\gamma}{n}\widehat{\Phi}^\top\widehat{\Phi})^j\Phi^\top\widehat{y} \ = \ (\widehat{\Phi}^\top\widehat{\Phi})^{-1}\widehat{\Phi}^\top\widehat{y}$$

Small t

$$\widehat{w}_t = (I - (I - \frac{\gamma}{n}\widehat{\Phi}^\top\widehat{\Phi}))^t(\widehat{\Phi}^\top\widehat{\Phi})^{-1}\Phi^\top\widehat{y} \underset{t=1}{\propto} \Phi^\top\widehat{y}$$

compare to

$$\widehat{w}_\lambda = (\widehat{\Phi}^\top\widehat{\Phi} + \lambda n I)^{-1}\Phi^\top\widehat{y} \underset{\lambda \text{ large}}{\propto} \Phi^\top\widehat{y}$$

Statistics

Theorem (Bauer, Pereverzev, R. '07)

If $p = \infty$, $\|\Phi(x)\|, |y| \leq 1$ *a.s. and* $t = \sqrt{n}$

$$\mathbb{E}[(\langle \Phi(x), \widehat{w}_t \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \frac{1}{\sqrt{n}}$$

Statistics

Theorem (Bauer, Pereverzev, R. '07)

If $p = \infty$, $\|\Phi(x)\|, |y| \leq 1$ *a.s. and* $t = \sqrt{n}$

$$\mathbb{E}[(\langle \Phi(x), \widehat{w}_t \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \frac{1}{\sqrt{n}}$$

Proof

$$\forall t > 1, \qquad \mathbb{E}[(\langle \Phi(x), \widehat{w}_t \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \frac{t}{n} + \frac{1}{t}$$

Remarks:

- Same bound as Tikhonov regularization.
- Known as iterative regularization (1955), early stopping (1985), implicit regularization (2018).

Computational regularization

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

Regularization by stochastic optimization

control statistics and time at once

What about other optimization methods?

Beyond gradient descent

What about other optimization methods?

- Accelerated methods (Conjugate Gradient, Nesterov, Heavyball).
- **Stochastic methods (SGD)**.

SGD with minibatches

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} \big(\langle \widehat{w}_t, \Phi(x_{j_i}) \rangle - y_{j_i}\big) \Phi(x_{j_i})$$

- b minibatch size ($b = 1$ SGD, $b = n$ GD)
- $t = \lceil \frac{n}{b} \rceil$ one pass

Statistics+Optimization

Theorem (Lin R. '16)

If $p = \infty$, $\|\Phi(x)\|, |y| \leq 1$ *a.s. if*

1. $b = 1$, $\gamma_t \simeq \frac{1}{\sqrt{n}}$, *and* $t = n$ *iterations* (1 *pass over the data*);
2. $b = \sqrt{n}$, $\gamma_t \simeq 1$, *and* $t = \sqrt{n}$ *iterations* (1 *pass over the data*);
3. $b = n$, $\gamma_t \simeq 1$, *and* $t = \sqrt{n}$ *iterations* ($\sqrt{n}$ *passes over the data*);

then,

$$\mathbb{E}[(\langle \Phi(x), \widehat{w}_t \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \frac{1}{\sqrt{n}}.$$

Statistics+Optimization

Theorem (Lin R. '16)

If $p = \infty$, $\|\Phi(x)\|, |y| \leq 1$ *a.s. if*

1. $b = 1$, $\gamma_t \simeq \frac{1}{\sqrt{n}}$, *and* $t = n$ *iterations (*1 *pass over the data);*
2. $b = \sqrt{n}$, $\gamma_t \simeq 1$, *and* $t = \sqrt{n}$ *iterations (*1 *pass over the data);*
3. $b = n$, $\gamma_t \simeq 1$, *and* $t = \sqrt{n}$ *iterations (*$\sqrt{n}$ *passes over the data);*

then,

$$\mathbb{E}[(\langle \Phi(x), \widehat{w}_t \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \frac{1}{\sqrt{n}}.$$

Proof

$$\forall \gamma > 0, t > 1, \qquad \mathbb{E}[(\langle \Phi(x), \widehat{w}_t \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \ \lesssim \frac{1}{\gamma t} + \frac{1}{\sqrt{n}} \left(\frac{\gamma t}{\sqrt{n}} \right)^2 + \frac{\gamma}{b} \left(1 + \frac{\gamma t}{\sqrt{n}} \right)$$

Other Flavors of SGD

- Cyclic, reshuffle SGD (R. Villa, '15)
- SGD with averaging (Bach et al. 14-. . . , Pillaud, Rudi, Bach '18)- larger step-size.
- Averaging as regularization (Neu, R. '18)

Computational regularization

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

Regularization by stochastic optimization

improved time while keeping optimal statistical error

What about memory costs?

Outline

Tackling memory with random projections

- Sketching & random features
- **Nyström methods & subsampling**

Subsampling aka Nyström methods

Consider

$$\overline{x}_1, \dots, \overline{x}_M \subset x_1, \dots, x_n$$

and

$$w_M = \sum_{j=1}^{M} \Phi(\overline{x}_j) c_j,$$

a random projection on a subsapace.

Connections to

- Nyström methds
- Galerkin methods
- Column subsampling

Linear algebra perspective

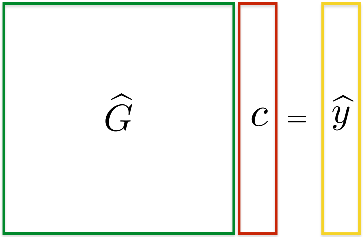
RandNLA=randomized numerical linear algebra (Halko et al. '11)

Back to ridge regression.

$$\langle \widehat{w}_\lambda, \Phi(x) \rangle = \sum_{i=1}^{n} \langle \Phi(x_i), \Phi(x) \rangle \, c^i$$

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

Linear System



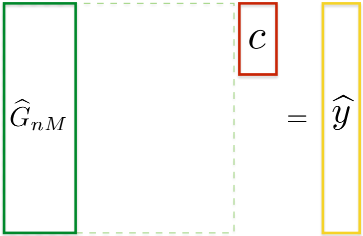
Nÿstrom/Column subsampling

Take $\overline{x}_1 \ldots, \overline{x}_M \subset x_1, \ldots, x_n, \quad M < n$.

$$\langle \widehat{w}_{\lambda,M}, \Phi(x) \rangle = \sum_{i=1}^{M} \langle \Phi(\overline{x}_i), \Phi(x) \rangle \, c^i$$

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

Linear System



Statistical guarantees

Theorem (Rudi, Camoriano, R. '16)

If $p = \infty$, $\|\Phi(x)\|, |y| \leq 1$ *a.s. if*

$$\lambda = 1/\sqrt{n}, \qquad M = \sqrt{n}$$

then,

$$\mathbb{E}[(\langle \Phi(x), \widehat{w}_{\lambda,M} \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \frac{1}{\sqrt{n}}$$

Remarks:

- Same bound again...
- Improve previous bounds (Bach et al. '12, Alaoui, Mahoney '14)
- Regularization by projection!

Computational regularization

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

Regularization by projection

control statistics, time and memory costs at once

Can we improve computational costs?

Outline

Preconditioning

Idea: define equivalent linear system with better condition number

Preconditioning

Idea: define equivalent linear system with better condition number

Preconditioning

$$(\widehat{G} + \lambda n I)c = \hat{y} \quad \mapsto \quad B^\top(\widehat{G} + \lambda n I)B\beta = B^\top \hat{y}, \quad c = B\beta.$$

Ideally $BB^\top = (\widehat{G} + \lambda n I)^{-1}$, so that

$$t = O(1/\lambda) \quad \mapsto \quad t = O(1)!$$

(Fasshauer et al '12, Avron et al '16, Cutajat '16, Ma, Belkin '17)

Baby FALKON

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

Preconditioning

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

Baby FALKON

$$\langle \widehat{w}_{\lambda,M,t}, \Phi(x)\rangle = \sum_{i=1}^{M} \langle \Phi(\widetilde{x}_i), \Phi(x)\rangle\, c^i \qquad c_t = B\beta_t$$

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

FALKON

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

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

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

Some Theory

Theorem (Rudi, Carratino, R. '17)

If $p = \infty$, $\|\Phi(x)\|, |y| \leq 1$ *a.s. if*

$$\lambda = 1/\sqrt{n}, \qquad M = \sqrt{n}, \qquad t = \log n$$

then

$$\mathbb{E}[(\langle \Phi(x), \widehat{w}_{\lambda,M,t} \rangle - \langle \Phi(x), w_* \rangle)^2] \lesssim \frac{1}{\sqrt{n}}$$

Remarks:

- Same bound again... improved time cost!
- Improved results by considering adaptive sampling.

Computational regularization

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

Maybe optimal?

Some experiments

	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%	-

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			HIGGS	
	c-err	AUC	Time(m)	AUC	Time(h)
FALKON	**19.6%**	0.877	**4**	0.825	**3**
EigenPro	19.8%	-	$6^{\wr}$	-	-
SVM	26.4%	-	9^{*}	-	-
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**	$78^{\dagger}$

Table: SUSY and HIGGS Datasets. Time obtained working on : $\dagger$ = cluster with 512 GB of RAM and IBM POWER8 12-core processor, $\wr$ = single machine with two Intel Xeon E5-2620, one Nvidia GTX Titan X GPU and 128GB of RAM, $\ddagger$ = single machine, $*$ = 14 workers.

Image classification

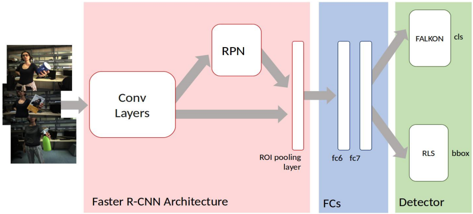
$$f(x) = \langle w, \Phi(x) \rangle \,, \quad x \mapsto \underbrace{\Phi_L}_{\text{Kernel representation}} \circ \underbrace{\Phi_{L-1} \cdots \circ \Phi_1(x)}_{\text{Convolutional}}$$

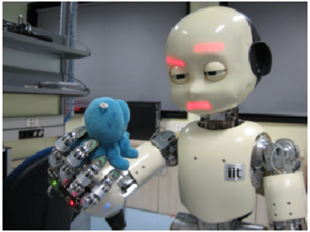
Imagenet

	Top-1 class error
FALKON + I-v3 feat.	**22.1%**
Inception-v3	21.2%
Inception-v2	23.4%
BN-Inception	25.2%
BN-GoogLeNet	26.8%
GoogLeNet	29.0%

Table: Single crop experimental results on the validation set of ILSVRC 2012.

Object detection





Method	mAP [%]	Train Time
Faster R-CNN	51,9	∼25 min
FALKON + Full Bootstrap (∼ 1K×1000)	51,5	∼8 min
FALKON + Random BKG (0 × 7000)	47,7	∼25 sec

Method	Train Time	mAP [%]	soda bottle	mug	pencil case	ring binder	wallet	flower	book	body lotion	hair clip	sprayer
Faster R-CNN Fine-tuning	∼40 min	49,7	63,2	68,4	23,3	29,6	49,9	66,1	35,3	56,2	60,2	45,8
FALKON + Random BKG (0×6000)	**∼25 sec**	40,5	57,7	67,9	17,5	23,1	23,8	59,5	26,8	39,6	48,5	40,5
FALKON + Mini Bootstrap (4×2500)	∼40 sec	48,1	63,1	67,2	18,4	25,7	47,4	70,3	36,1	52,3	58,8	41,5
FALKON + Mini Bootstrap (10×1500)	∼50 sec	**51,3**	64,7	71,2	27,2	31,7	56,9	69,4	39,6	54,0	60,7	37,1

Summing up

- Computational regularization for efficient learning.
- Faster GP/Kernel solver ever.

Looking ahead

- Other loss functions, norms, learning problems ....
- Parallelization.
- Non convex problems.
- Optimal complexity.

check papers on `arxiv.org`