Learning with SGD: bridging theory and practice

Lorenzo Rosasco
Universitá di Genova
Massachusetts Institute of Technology - Istituto Italiano di Tecnologia

joint work with:
S. Villa (Universitá di Genova), J. Lin (Zhejiang University),
G. Neu (Pompeu Fabra), N Mücke (University Stuttgart)

April 3, 2019

Outline

Warm-up: SGD in theory and in practice

Least squares learning with SGD

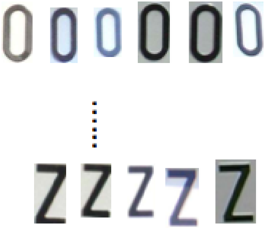
Multipass SGD+all the tricks

Machine learning applications

Texts

Subject	Date	Time	Body	Spam?
I has the viagra for you	03/12/1992	12:23 pm	Hi! I noticed that you are a software engineer so here's the pleasure you were looking for...	Yes
Important business	05/29/1995	01:24 pm	Give me your account number and you'll be rich. I'm totally serial	Yes
Business Plan	05/23/1996	07:19 pm	As per our conversation, here's the business plan for our new venture Warm regards...	No
Job Opportunity	02/29/1998	08:19 am	Hi !I am trying to fill a position for a PHP ...	Yes
[A few thousand rows ommitted]				
Call mom	05/23/2000	02:14 pm	Call mom. She's been trying to reach you for a few days now	No

Images



Data: $(x_1, y_1), \ldots, (x_n, y_n)$

Note: $x_i \in \mathbb{R}^d$ with d, n potentially *huge*!

Accuracy vs efficiency

Stochastic gradient descent - SGD

Stochastic optimization and SGD

Problem

Solve

$$\min_{w \in \mathcal{H}} \mathbb{E}_Z[\ell(w, Z)]$$

given $z_1, \ldots, z_n$ i.i.d.

Stochastic optimization and SGD

Problem

Solve

$$\min_{w \in \mathcal{H}} \mathbb{E}_Z[\ell(w, Z)]$$

given $z_1, \dots, z_n$ i.i.d.

SGD

$$\hat{w}_{t+1} = \hat{w}_t - \eta_t \nabla \ell(\hat{w}_t, z_t), \qquad t = 0, 1, \dots, n$$

- $\mathbb{E}_{Z_t} \nabla \ell(w, Z_t) = \nabla \mathbb{E}_{Z_t}[\ell(w, Z_t)])$ hence the name! (albeit it is not a descent method...)

[Robbins Munro '51...]

SGD in theory

Let

$$\overline{w}_n = \frac{1}{n+1} \sum_{t=0}^{n} \hat{w}_t \qquad\qquad w^\dagger = \arg\min_{w \in \mathcal{H}} \mathbb{E}_Z[\ell(w, Z)]$$

Then for L convex

$$\eta_t \simeq 1/\sqrt{n} \qquad \Rightarrow \qquad L(\overline{w}_n) - L(w^\dagger) = O(1/\sqrt{n})$$

Note: One pass SGD: data points are used once, iterations are conditionally independent.

[Nemirovski, Yudin '83, Agarwal et al. '12]

SGD in practice

In practice:

- multiple passes $t > n$
- data-adaptive step-size
- mini-batching
- different forms of averaging.

Implicit regularization

Outline

Warm-up: SGD in theory and in practice

Least squares learning with SGD

Multipass SGD+all the tricks

Least squares learning

$Z = (X, Y) \sim \rho$ on $\mathcal{X} \times \mathbb{R}$, $\mathcal{X}$ real separable Hilbert space (linear/functional regression RKHS).

Problem:
Solve

$$\min_{w \in \mathcal{X}} L(w) \qquad\qquad L(w) = \frac{1}{2}\, \mathbb{E}[(Y - \langle w, X \rangle)^2]$$

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

Least squares learning

$Z = (X, Y) \sim \rho$ on $\mathcal{X} \times \mathbb{R}$, $\mathcal{X}$ real separable Hilbert space (linear/functional regression RKHS).

Problem:
Solve

$$\min_{w \in \mathcal{X}} L(w) \qquad\qquad L(w) = \frac{1}{2}\,\mathbb{E}[(Y - \langle w, X\rangle)^2]$$

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

Least squares optimality conditions

$$\Sigma w = g, \qquad\qquad \Sigma = \mathbb{E}[X \otimes X], \quad h = \mathbb{E}[XY].$$

Least squares learning

$Z = (X, Y) \sim \rho$ on $\mathcal{X} \times \mathbb{R}$, $\mathcal{X}$ real separable Hilbert space (linear/functional regression RKHS).

Problem:
Solve

$$\min_{w \in \mathcal{X}} L(w) \qquad\qquad L(w) = \frac{1}{2}\mathbb{E}[(Y - \langle w, X \rangle)^2]$$

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

Least squares optimality conditions

$$\Sigma w = g, \qquad\qquad \Sigma = \mathbb{E}[X \otimes X], \quad h = \mathbb{E}[XY].$$

Ill-posedness

- **$\mathcal{X}$ infinite dimensional, Σ compact $\Rightarrow$ problem is ill-posed.**
- if $\mathcal{X}$ is finite dimensional it is well posed but potentially ill-conditioned.

Minimal norm solution

Moore-Penrose solution:

$$w^\dagger = \underset{w \in \mathcal{X}}{\arg\min} \, \|w\| \,, \qquad \text{subj. to } \; \Sigma w = g.$$

Minimal norm solution

Moore-Penrose solution:

$$w^\dagger = \arg\min_{w \in \mathcal{X}} \|w\|, \qquad \text{subj. to } \Sigma w = g.$$

Regularization

- Looking for a minimal norm solution $=$ bias in the estimation process.
- Minimal norm solution can be unstable to noise/sampling $\rightarrow$ *regularization*.

Multi-pass SGD

$$\widehat{w}_{t+1} = \widehat{w}_t - \eta_t \left(x_{i_t} (\langle \widehat{w}_t, x_{i_t} \rangle - y_{i_t}) \right), \quad t = 0, \dots T$$

Algorithmic choices

- i_t deterministic or stochastic selection (with/without replacement);
- step-size η_t;
- stopping time T ($T > n$ multiple *"passes"*).

No explicit penalties or constraints.

SOA: Incremental gradient for ERM

$$\widehat{w}_{t+1} = \widehat{w}_t - \eta_t \left(x_{i_t} (\langle \widehat{w}_t, x_{i_t} \rangle - y_{i_t}) \right), \quad t = 0, \dots T$$

Empirical risk minimization (ERM)

$$\min_{w \in \mathcal{X}} \widehat{L}(w) \qquad \widehat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, x_i \rangle)^2$$

Then

$$\eta_t \simeq 1/n^a \qquad \Rightarrow \qquad \widehat{L}(\overline{w}_n) - \min \widehat{L}(w) = O(1/\sqrt{n})$$

[Bertsekas '97]

We are interested in the expect error L.

Learning with cyclic SGD

Recall $\Sigma w^\dagger = g$.

Assumption **A)** $\left\|\Sigma^{-\alpha} w^\dagger\right\| \leq R$, $\alpha > 0$.

- infinite dimensional extension of KL condition [Garrigos, R., Villa '18].

Learning with cyclic SGD

Recall $\Sigma w^\dagger = g$.

Assumption **A)** $\quad \left\|\Sigma^{-\alpha} w^\dagger\right\| \leq R$, $\alpha > 0$.

• infinite dimensional extension of KL condition [Garrigos, R., Villa '18].

Theorem (R. Villa '15)

Assume $\|x\| \leq 1$ *and* $|y| \leq 1$ *and A). If* $\eta = O(1/n)$ *then for* $t \in \mathbb{N}$ *whp*

$$\left\|\hat{w}_t - w^\dagger\right\|^2 \lesssim \frac{t^2}{n} + \frac{1}{t^{2\alpha}},$$

so that for $T \simeq n^{\frac{1}{2(\alpha+1)}}$ *whp*

$$\left\|\hat{w}_T - w^\dagger\right\|^2 \lesssim n^{-\frac{\alpha}{\alpha+1}}.$$

Proof strategy

Samples reused in multiple iterations, hence no conditional independence.

Let

$$w_{t+1} = w_t - \eta \nabla L(w_t)$$

$$w_t \longrightarrow w_{t+1} \quad \longrightarrow \dots \quad \longrightarrow \quad \longrightarrow w^\dagger$$

Proof strategy

Samples reused in multiple iterations, hence no conditional independence.

Let

$$w_{t+1} = w_t - \eta \nabla L(w_t)$$

$$w_t \longrightarrow w_{t+1} \quad \longrightarrow \ldots \quad \longrightarrow \quad \longrightarrow w^\dagger$$

$$\widehat{w}_t \longrightarrow \widehat{w}_{t+1} \quad \longrightarrow \ldots$$

Proof strategy

Samples reused in multiple iterations, hence no conditional independence.

Let

$$w_{t+1} = w_t - \eta \nabla L(w_t)$$

$$w_t \longrightarrow w_{t+1} \quad \longrightarrow \dots \quad \longrightarrow \quad \longrightarrow w^\dagger$$

$$\widehat{w}_t \longrightarrow \widehat{w}_{t+1} \quad \longrightarrow \dots$$

$$\searrow$$

$$\searrow$$

$$\arg\min_{\mathcal{X}} \widehat{L}$$

Elements of the proof

Optimization/Bias

$$w_t = (I - \eta\Sigma)w_t + \eta h = \eta \sum_{j=0}^{t-1} (I - \eta\Sigma)^j h \qquad\qquad w_t - w^\dagger = (I - \eta\Sigma)^t w^\dagger$$

Stability/Variance

$$\widehat{w}_{t+1} = \underbrace{(I - \eta\widehat{\Sigma})\widehat{w}_t + \eta\widehat{h}}_{\text{batch GD}} + \underbrace{\eta^2 \widehat{e}_t}_{\text{"noise"}}, \qquad\qquad \widehat{e}_t = \widehat{A}\widehat{w}_t - \widehat{b}$$

with

$$\widehat{A} = \frac{1}{n^2} \sum_{k=2}^{n} \prod_{i=k+1}^{n} \left(I - \frac{1}{n} x_i \otimes x_i\right) x_k \otimes x_k \sum_{j=1}^{k-1} x_k \otimes x_j$$

random variable with **martingale** structure...

Remarks

- No averaging "deterministic" multipass SGD converges and iterates rates are optimal.
- The obtained results match those for regularized ERM with $\lambda = 1/t$,

$$\hat{w}_\lambda = \min_{w \in \mathcal{X}} \frac{1}{n} \sum_{i=1} (y_i - \langle w, x_i \rangle)^2 + \lambda \left\| w \right\|^2$$

Remarks

- No averaging "deterministic" multipass SGD converges and iterates rates are optimal.
- The obtained results match those for regularized ERM with $\lambda = 1/t$,

$$\hat{w}_\lambda = \min_{w \in \mathcal{X}} \frac{1}{n} \sum_{i=1} (y_i - \langle w, x_i \rangle)^2 + \lambda \|w\|^2$$

- SGD performs implicit/iterative regularization: it converges to the minimal norm solution;
- the number of iterations parameterize regularization;

Remarks

- No averaging "deterministic" multipass SGD converges and iterates rates are optimal.
- The obtained results match those for regularized ERM with $\lambda = 1/t$,

$$\hat{w}_\lambda = \min_{w \in \mathcal{X}} \frac{1}{n} \sum_{i=1} (y_i - \langle w, x_i \rangle)^2 + \lambda \|w\|^2$$

- SGD performs implicit/iterative regularization: it converges to the minimal norm solution;
- the number of iterations parameterize regularization;
- same rates with data driven tuning (e.g. cv, Lepskii [R. Perverzev, De Vito '07, Caponnetto, Yao '06]).

Missing: sharp value expected loss bounds.

Outline

Warm-up: SGD in theory and in practice

Least squares learning with SGD

Multipass SGD+all the tricks

The "stochastic" SGD

$$\widehat{w}_{t+1} = \widehat{w}_t - \eta x_{i_t}(\langle \widehat{w}_t, x_{i_t} \rangle - y_{i_t}), \quad t = 0, \dots T$$

$(i_t)_t$ chosen **uniformly at random with replacement**

Multipass SGD: worst case

Theorem (Lin, R. '17)

Assume $\|x\| \leq 1$ *and* $|y| \leq 1$ *then for all* η *and* t,

$$\mathbb{E}\, L(\widehat{w}_t) - L(w^\dagger) \lesssim \frac{1}{\sqrt{n}} \left(\frac{\eta t}{\sqrt{n}} \right)^2 + \eta \left(1 \vee \frac{\eta t}{\sqrt{n}} \right) + \frac{1}{\eta t}.$$

If

- $T \simeq n^{3/2}$ *(*$\sqrt{n}$ *passes),* $\eta \simeq \frac{1}{n}$, *or*
- $T \simeq n$ *(*1 *pass),* $\eta \simeq \frac{1}{\sqrt{n}}$,

then,

$$\mathbb{E}\, L(\widehat{w}_T) - L(w^\dagger) \lesssim \frac{1}{\sqrt{n}}.$$

Remarks

- No averaging multipass SGD converges and learning rates are optimal- same as ERM;

Remarks

- No averaging multipass SGD converges and learning rates are optimal- same as ERM;
- the product of the number of iterations parameterize regularization;
- implicit/iterative regularization & regularization;
- similar results for the iterats, SGD converges to the minimal norm solution.

Remarks

- No averaging multipass SGD converges and learning rates are optimal- same as ERM;
- the product of the number of iterations parameterize regularization;
- implicit/iterative regularization & regularization;
- similar results for the iterats, SGD converges to the minimal norm solution.

What about faster rates?

Beyond the worst case

Least squares optimality conditions

$$\Sigma w^{\dagger} = g,$$

Beyond the worst case

Least squares optimality conditions

$$\Sigma w^{\dagger} = g,$$

Assumptions

- **A** $\left\|\Sigma^{-\alpha} w^{\dagger}\right\| \leq R$, $\alpha > 0$

- **Capacity** $\sigma_i(\Sigma) \sim i^{-\frac{1}{\gamma}}, \quad \gamma \in (0, 1]$

• Reduces to worst case for $\alpha = 0, \quad \gamma = 1.$

Multipass SGD: fast rates

Theorem (Lin, R. '17)

Assume $\|x\| \leq 1$, $|y| \leq 1$ *and* $A), C)$ *hold. Then, for all* η *and* t,

$$\mathbb{E}\, L(\widehat{w}_t) - L(w^\dagger) \lesssim \left(\frac{1}{\eta t}\right)^{2\alpha+1} + \frac{1}{n^{\frac{2\alpha+1}{2\alpha+1+\gamma}}}\left(\frac{\eta t}{n^{\frac{1}{2\alpha+1+\gamma}}}\right)^2 + \eta\left(1 \vee \frac{\eta t}{n^{\frac{1}{2\alpha+1+\gamma}}}\right).$$

If

- $T \simeq n^{\frac{1}{2\alpha+1+\gamma}+1}$ ($n^{\frac{1}{2\alpha+1+\gamma}}$ *passes*), $\eta \simeq \frac{1}{n}$,
- $T \simeq n$ (1 *pass*), $\eta \simeq n^{-\frac{2\alpha+1}{2\alpha+1+\gamma}}$,

then,

$$\mathbb{E}\, L(\widehat{w}_{T_n}) - L(w^\dagger) \lesssim n^{-\frac{2\alpha+1}{2\alpha+1+\gamma}}$$

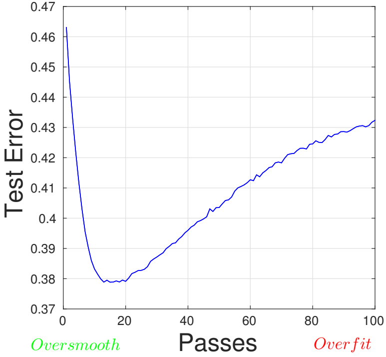
Remarks

- No averaging multipass SGD converges with fast learning rates- same as ERM;
- implicit/iterative regularization;
- optimal parameters choice depends on uknowns;
- same rates with cross validation/Lepskii method [R. Perverzev, De Vito '07, Caponnetto, Yao '06]).

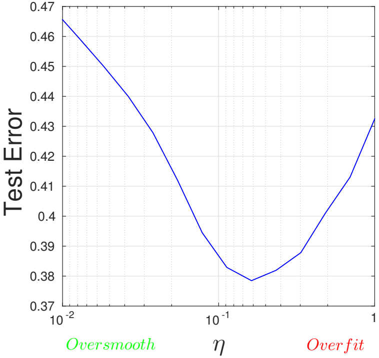
SGM in pratice

Model selection on $\#$ of passes and/or η!

Fixed η



1 pass



Elements of the proof

Let

$$\underbrace{w_t = \eta \sum_{j=0}^{t-1} (I - \eta\Sigma)^j h,}_{\text{Population GD}} \qquad \underbrace{\tilde{w}_t = \eta \sum_{j=0}^{t-1} (I - \eta\widehat{\Sigma})^j \widehat{h}}_{\text{Batch GD}}$$

Optimization/Bias

$$w^\dagger - w_t = (I - \eta\Sigma)^t w^\dagger$$

Stability/Sample variance

$$w_t - \widetilde{w}_t = \eta \sum_{j=0}^{t-1} (I - \eta\Sigma)^j h - \eta \sum_{j=0}^{t-1} (I - \eta\widehat{\Sigma})^j \widehat{h}$$

Stability/Computational variance

$$\widetilde{w}_t - \widehat{w}_t, \qquad \widehat{w}_t = \mathbb{E}\, \widetilde{w}_t$$

SGD in practice

In practice:

- multiple passes $t > n$, ✓
- data-adaptive step-size, ✓
- mini-batching
- different forms of averaging.

The "stochastic" SGD

$$\widehat{w}_{t+1} = \widehat{w}_t - \eta \frac{1}{b} \sum_{j=b(t-1)}^{bt} x_{i_j}(\langle \widehat{w}_t, x_{i_j} \rangle - y_{i_j}), \quad t = 0, \dots T$$

Algorithmic choices

- b mini-batch size
- $\lceil bt/n \rceil$ number of passes
- $(i_t)_t$ chosen **uniformly at random with replacement**

Mini-batch SGD worst case

Theorem (Lin, R. '17)

Assume $\|x\| \leq 1$ *and* $|y| \leq 1$ *for all* η *and* t,

$$\mathbb{E}\, L(\widehat{w}_t) - L(w^\dagger) \lesssim \frac{1}{\eta t} + \frac{1}{\sqrt{n}} \left(\frac{\eta t}{\sqrt{n}} \right)^2 + \frac{\eta}{b} \left(1 + \frac{\eta t}{\sqrt{n}} \right).$$

If

- $b \simeq 1$, $T \simeq n$ *(*1 *pass),* $\eta \simeq \frac{1}{\sqrt{n}}$,
- $b \simeq \sqrt{n}$, $T \simeq \sqrt{n}$ *(*1 *pass),* $\eta \simeq 1$,
- $b > \sqrt{n}$, $T > \sqrt{n}$ *(*> 1 *pass),* $\eta \simeq 1$,

then,

$$\mathbb{E}\, L(\widehat{w}_T) - L(w^\dagger) \lesssim \frac{1}{\sqrt{n}}.$$

Remarks

- Mini-batching allows larger step-size.
- There's a critical mini-batch size ($b = \sqrt{n}$) after which there's no gain.
- The mini-batch size controls the SGD learning behavior together with step-size and $\#$ of iterations.

Faster rates?

Mini-batch SGD fast rates

Theorem (Lin, R. '17)

Assume $\|x\| \leq 1$, $|y| \leq 1$ *and* $A), C)$ *hold. Then, for all* η *and* t,

$$\mathbb{E}\, L(\widehat{w}_t) - L(w^\dagger) \lesssim \left(\frac{1}{\eta t}\right)^{2\alpha+1} + \frac{1}{n^{\frac{2\alpha+1}{2\alpha+1+\gamma}}} \left(\frac{\eta t}{n^{\frac{1}{2\alpha+1+\gamma}}}\right)^2 + \frac{\eta}{b}\left(1 \vee \frac{\eta t}{n^{\frac{1}{2\alpha+1+\gamma}}}\right).$$

If

- $b \simeq 1$, $T \simeq n$, $\eta \simeq n^{-\frac{2\alpha+1}{2\alpha+1+\gamma}}$
- $b \simeq n^{\frac{2\alpha+1}{2\alpha+1+\gamma}}$, $T \simeq n^{\frac{1}{2\alpha+1+\gamma}}$, $\eta \simeq 1$
- $b \simeq n$, $T \simeq n^{\frac{1}{2\alpha+1+\gamma}}$, $\eta \simeq 1$,

then,

$$\mathbb{E}\, L(\widehat{w}_T) - L(w^\dagger) \lesssim n^{-\frac{2\alpha+1}{2\alpha+1+\gamma}}.$$

Remarks

- Different way to control the properties of SGD choosing b, η, T.

- Again a critical mini-batch size, now depending on the regularity of the problem.

- Analogous results hold for data driven tuning (e.g. cv, Lepskii [R. Perverzev, De Vito '07, Caponnetto, Yao '06]).

Missing: Averaging leads to larger step-sizes for one pass [Bach, Moulines '13, Dieuleveut, Bach'16] ... but also slower learning rates in some regimes (saturation).

Tail-averaged SGM

$$\overline{w}_L = \frac{1}{T-S}\sum_{t=S+1}^{T} \widehat{w}_t$$

Algorithmic choices

- $S = 0$ uniform averaging,
- $L = T - S$ tail lenght.

An insight from GD

Population GD: $w_{t+1} = (I - \eta\Sigma)w_t + h$,

$$w_t - w^\dagger = (I - \eta\Sigma)^t w^\dagger \qquad O\left(\frac{1}{t^{2\alpha+1}}\right)$$

if $\left\|\Sigma^{-\alpha} w^\dagger\right\| \leq R$, $\alpha > 0$.

An insight from GD

Population GD: $w_{t+1} = (I - \eta\Sigma)w_t + h$,

$$w_t - w^\dagger = (I - \eta\Sigma)^t w^\dagger \qquad O\left(\frac{1}{t^{2\alpha+1}}\right)$$

if $\left\|\Sigma^{-\alpha} w^\dagger\right\| \leq R$, $\alpha > 0$.

Tail-averaged population GD: $\tilde{w}_L = \frac{1}{T-S}\sum_{t=S+1}^{T} w_t$,

$$\tilde{w}_L - w^\dagger \approx \frac{(I - \eta\Sigma)^{S+1}}{T} w^\dagger,$$

the rate is is $O\left(\frac{1}{t^{2\alpha+1}}\right)$ if $S \propto T$ and at most $1/T$ for $S = 0$ [Mücke, Neu, R. '19].

Mini-batch SGD fast rates

Theorem (Mücke, Neu, R. '19)

Assume $\|x\| \leq 1$, $|y| \leq 1$ *and* $A), C)$ *hold. Then, for all* η *and* $L = t - S$, *and* $S = 0$, $\alpha \leq 1/2$ *or* $S \propto T$, $\alpha > 0$

$$\mathbb{E}\, L(\overline{w}_L) - L(w^\dagger) \lesssim \frac{1}{(\eta L)^{2\alpha+1}} + \frac{(\eta L)^\gamma}{n} + \frac{\eta}{b(\eta L)^{(1-\alpha)}}$$

If

- $b \simeq 1$, $L \simeq n$, $\eta \simeq n^{-\frac{2\alpha+\gamma}{2\alpha+1+\gamma}}$
- $b \simeq n^{\frac{2\alpha+\gamma}{2\alpha+1+\gamma}}$, $L \simeq n^{\frac{1}{2\alpha+1+\gamma}}$, $\eta \simeq 1$
- $b \simeq n$, $L \simeq n^{\frac{1}{2\alpha+1+\gamma}}$, $\eta \simeq 1$.

then,

$$\mathbb{E}\, L(\overline{w}_L) - L(w^\dagger) \lesssim n^{-\frac{2\alpha+1}{2\alpha+1+\gamma}}$$

Remarks

- For one pass $\alpha \le 1/2$, we recover the results of [Dieuleveut, Bach '16] for uniform averaging $S = 0$.
- We extend these results to $\alpha > 1/2$ via tail averaging.
- Compared to [Lin, R. '17] we obtain a smaller critical minibatch size

$$b_n \simeq n^{\frac{2\alpha+\gamma}{2\alpha+1+\gamma}} \qquad \text{instead of} \qquad b_n \simeq n^{\frac{2\alpha+1}{2\alpha+1+\gamma}}$$

- Nonparametric analogue of the results in [Jain et al. '18].
- The proof combines ideas from [Lin, R. '17] and [Pillaud et al '18]

Summing up

- Learning properties of practical SGD & implicit regularization
- Further: combine random projectons with SGD [Carratino, Rudi, R. '18]
- Further: consider different learning regimes [Pillaud, Rudi, Bach '18]
- TBD: other losses, other norms, other functions (deep nets?)

All papers on `arxiv.org`: [Villa, Rosasco '15, Lin, Rosasco' 17, Mücke, Neu, Rosasco '19]

Shameless plug:

European Research Council
Executive Agency

Multiple openings for post-docs/PhD positions!

@lrntzrsc

→ **Launching MaLGa: Machine Learning Genova Center!**