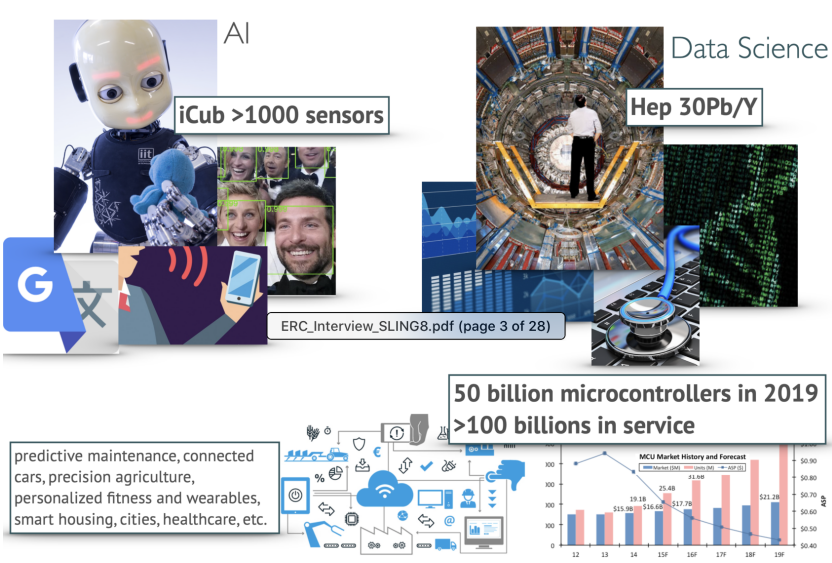
UniGe | MaLGa

A guided tour of machine learning (theory)

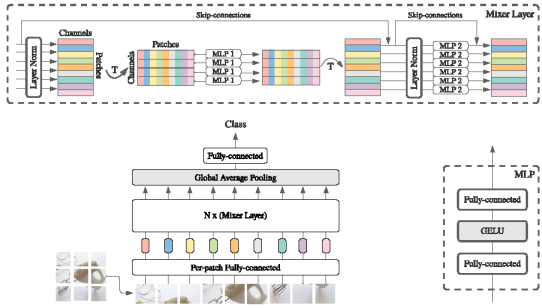
Lorenzo Rosasco
MaLGa, Università degli Studi di Genova, MIT, IIT

RoMaDS Colloquim, University of Rome Tor Vergata, May 9th 2022

AI everywhere (literally...)



The state of affairs



Rethinking machine learning:

- with statistical mechanics
- with information theory
- with tropical geometry
- ...

Outline

Data driven modeling paradigm

Statistical learning, bias and variance

From statistics to optimization

A theory crisis?

The basic picture

$$(x_i, y_i)_{i=1}^n \quad \mapsto \quad f : X \to Y$$

Fixing a model

$$w \in \mathbb{R}^p \;\mapsto\; f_w$$

Fixing a model

$$w \in \mathbb{R}^p \;\mapsto\; f_w$$

$$f_w(x) = \sum_{j=1}^{p} w^j \phi_j(x)$$

Fixing a model

$$w \in \mathbb{R}^p \;\mapsto\; f_w$$

$$f_w(x) = \sum_{j=1}^{p} w^j \phi_j(x)$$

$$f_w(x) = \sum_{j=1}^{p} \beta^j \sigma(a_j^\top x + \alpha_j),$$

Fitting (or overfitting??)

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} (f_w(x_i) - y_i)^2$$

Fitting (or overfitting??)

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} (f_w(x_i) - y_i)^2$$

Overparametrization

w has often millions of parameters...data are often (much) less!

Fitting (or overfitting??)

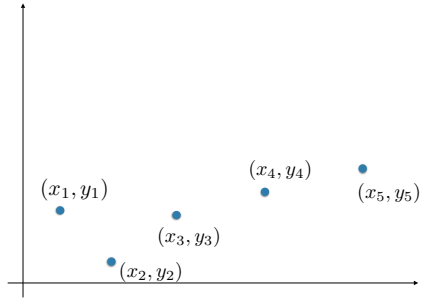
$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} (f_w(x_i) - y_i)^2$$

Overparametrization

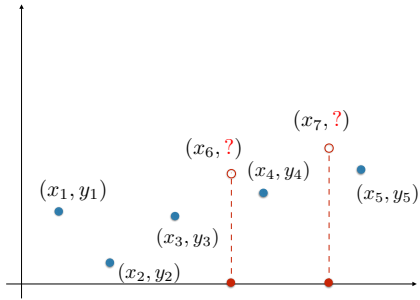
w has often millions of parameters...data are often (much) less!

> "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk"
>
> von Neumann

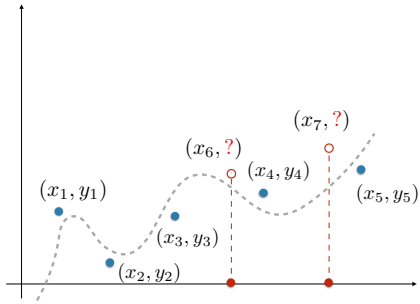
Learning is not (just) fitting, but prediction



Learning is not (just) fitting, but prediction



Learning is not (just) fitting, but prediction



Predictions from random and noisy samples

Learning pipeline

Model fitting (regularized)

$$\widehat{w}_\theta = \underset{\|w\| \leqslant \theta}{\operatorname{argmin}} \frac{3}{n} \sum_{i=1}^{n/3} (f_w(x_i) - y_i)^2$$

Learning pipeline

Model fitting (regularized)

$$\widehat{w}_\theta = \underset{\|w\| \leqslant \theta}{\operatorname{argmin}} \frac{3}{n} \sum_{i=1}^{n/3} (f_w(x_i) - y_i)^2$$

Model tuning

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{3}{n} \sum_{i=n/3+1}^{2n/3} (f_{\widehat{w}_\theta}(x_i) - y_i)^2$$

Learning pipeline

Model fitting (regularized)

$$\widehat{w}_\theta = \underset{\|w\| \leqslant \theta}{\operatorname{argmin}} \frac{3}{n} \sum_{i=1}^{n/3} (f_w(x_i) - y_i)^2$$

Model tuning

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{3}{n} \sum_{i=n/3+1}^{2n/3} (f_{\widehat{w}_\theta}(x_i) - y_i)^2$$

Model assessment

$$\frac{3}{n} \sum_{i=2n/3+1}^{n} (f_{\widehat{w}_{\widehat{\theta}\,\theta}}(x_i) - y_i)^2$$

Classic vs data driven modeling

- Paradigm shift in modeling, **driven by data** availability.
- Careful **pipeline** needed.
- **Theoretical** guidance needed.

ML theory

- Representation: ”Which model?”
- Generalization: ”How accurate is my model?”
- Optimization: ”How can I compute my model?”

Outline

Data driven modeling paradigm

Statistical learning, bias and variance

From statistics to optimization

A theory crisis?

Statistical machine learning

- $(X, Y) \sim P$ random variables in $\mathbb{R}^d \times \mathbb{R}$, and $(x_1, y_1), \ldots, (x_n, y_n) \sim P^n$.

Statistical machine learning

- $(X, Y) \sim P$ random variables in $\mathbb{R}^d \times \mathbb{R}$, and $(x_1, y_1), \ldots, (x_n, y_n) \sim P^n$.
- $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ loss function, e.g. $\ell(f(x), y) = (y - f(x))^2$.

Statistical machine learning

- $(X, Y) \sim P$ random variables in $\mathbb{R}^d \times \mathbb{R}$, and $(x_1, y_1), \ldots, (x_n, y_n) \sim P^n$.
- $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ loss function, e.g. $\ell(f(x), y) = (y - f(x))^2$.

Problem: minimize

$$L(f) = \mathbb{E}[\ell(f(X), Y)],$$

given only $(x_1, y_1), \ldots, (x_n, y_n) \sim P^n$.

ERM and its excess risk

$$\widehat{w}_\theta = \operatorname*{argmin}_{\|w\| \leqslant \theta} \widehat{L}(f_w), \qquad \widehat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

$$\widehat{f}_\theta = f_{\widehat{w}_\theta}$$

ERM and its excess risk

$$\widehat{w}_\theta = \operatorname*{argmin}_{\|w\| \leqslant \theta} \widehat{L}(f_w), \qquad \widehat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

$$\widehat{f}_\theta = f_{\widehat{w}_\theta}$$

Excess risk

$$L(\widehat{f}_\theta) - \min L(f)$$

Error decomposition

Population algorithm

$$f_\theta = f_{w_\theta}, \qquad w_\theta = \operatorname*{argmin}_{\|w\| \leqslant \theta} L(f_w)$$

Error decomposition

Population algorithm

$$f_\theta = f_{w_\theta}, \qquad w_\theta = \operatorname*{argmin}_{\|w\| \leqslant \theta} L(f_w)$$

$$L(\widehat{f}_\theta) - \min L(f) = \underbrace{L(\widehat{f}_\theta) - L(f_\theta)}_{\text{Estimation error}} + \underbrace{L(f_\theta) - \min L(f)}_{\text{Approximation error}}$$

Approximation error

Assume

$$|\ell(y, f(x)) - \ell(y, f(x))| \leqslant C_\ell |f(x) - f'(x)|$$

Lemma

Let $L(f_*) = \min L(f)$, *then*

$$L(f_\theta) - \min L(f) \leqslant C_\ell \min_{\|w\| \leqslant \theta} \|f_\theta - f_*\|_{L^1(P)}$$

Approximation error

Assume

$$|\ell(y, f(x)) - \ell(y, f(x))| \leqslant C_\ell |f(x) - f'(x)|$$

Lemma

Let $L(f_*) = \min L(f)$, *then*

$$L(f_\theta) - \min L(f) \leqslant C_\ell \min_{\|w\| \leqslant \theta} \|f_\theta - f_*\|_{L^1(P)}$$

Proof.

$$L(f_\theta) - L(f_*) = \min_{\|w\| \leqslant \theta} \int (\ell(f(x), y) - \ell(f_*(x), y)) dP(x, y) \leqslant C_\ell \min_{\|w\| \leqslant \theta} \int |f(x) - f_*(x)| dP(x, y)$$

□

[Cucker, Zhou '93]

Universality

A model is universal if for all f_*

$$\lim_{\theta \to \infty} \|f_\theta - f_*\|_{L^1(P)} = 0.$$

e.g. Kernel methods and neural nets.

[DeVore, Lorentz '93, Pinkus '99]

Smoohtness conditions

Assume

$$f_* \in \mathcal{H}_s,$$

for some smoothness class $\mathcal{H}_s$. e.g. the Sobolev space $W^{s,2}$.

Smoohtness conditions

Assume

$$f_* \in \mathcal{H}_s,$$

for some smoothness class $\mathcal{H}_s$. e.g. the Sobolev space $W^{s,2}$.

Approximation results ensure that

$$\min_{\|w\| \leqslant \theta} \|f_\theta - f_*\|_{L^1(P)} \lesssim a(\theta, s)$$

where $a(\theta, a)$ decays for θ incresing and rate depending on s, e.g. $\theta^{-s/d}$

[DeVore, Lorentz, '93]

Estimation error

Lemma

By definition of ERM, it holds

$$L(\widehat{f}_\theta) - L(f_\theta) \leqslant C_\ell \sup_{\|w\| \leqslant \theta} |\widehat{L}(f_w) - L(f_w)|$$

Estimation error

Lemma

By definition of ERM, it holds

$$L(\widehat{f}_\theta) - L(f_\theta) \leqslant C_\ell \sup_{\|w\| \leqslant \theta} |\widehat{L}(f_w) - L(f_w)|$$

Proof.

$$L(\widehat{f}_\theta) - L(f_\theta) = L(\widehat{f}_\theta) - \widehat{L}(\widehat{f}_\theta) + \underbrace{\widehat{L}(\widehat{f}_\theta) - \widehat{L}(f_\theta)}_{\leqslant 0} + \widehat{L}(f_\theta) - L(f_\theta)$$

□

[Vapnilk, Chervonenkis, '77, Gyorfi, Devroye, Lugosi, '96]

Capacity measures

Empirical process

$$\sup_{\|w\| \leqslant \theta} |\widehat{L}(f_w) - L(f_w)|$$

Capacity measures

Empirical process

$$\sup_{\|w\| \leqslant \theta} |\widehat{L}(f_w) - L(f_w)|$$

Lemma (Rademacher complexities)

If $\sigma_i \in \{\pm 1\}, P(1) = P(-1) = 1/2, i = 1, \dots, n$ *(Rademacher random variables), then*

$$\mathbb{E}\left[\sup_{\|w\| \leqslant \theta} |\widehat{L}(f_w) - L(f_w)|\right] \leqslant 2C_\ell \underbrace{\mathbb{E}\left[\frac{1}{n} \sup_{\|w\| \leqslant \theta} \sum_{i=1}^{n} \sigma_i f_w(x_i)\right]}_{\text{Rademacher complexity}},$$

Capacity measures for linear models

$$f_w = \sum_{j=1}^{\infty} w^j \phi_j$$

Capacity measures for linear models

$$f_w = \sum_{j=1}^{\infty} w^j \phi_j$$

Lemma

If

$$\sup_x |\sum_{j=1}^{\infty} \phi_j(x)|^2 \leqslant \kappa^2$$

then

$$\mathbb{E}\left[\frac{1}{n} \sup_{\|w\| \leqslant \theta} \sum_{i=1}^{n} \sigma_i f_w(x_i)\right] \leqslant \frac{\theta\kappa}{\sqrt{n}}$$

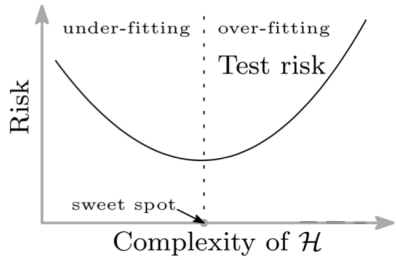
Results for nonlinear models can be similarly derived.

The bias-variance trade-off

$$L(\widehat{f}_\theta) - \min L(f) \lesssim \frac{\theta}{\sqrt{n}} + a(\theta, s)$$

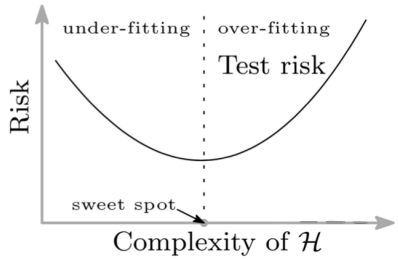
The bias-variance trade-off

$$L(\widehat{f}_\theta) - \min L(f) \lesssim \frac{\theta}{\sqrt{n}} + a(\theta, s)$$



The bias-variance trade-off

$$L(\widehat{f}_\theta) - \min L(f) \lesssim \frac{\theta}{\sqrt{n}} + a(\theta, s)$$



$$\theta_* = \theta(s, n) \quad \Longrightarrow \quad L(\widehat{f}_{\theta_*}) - \min L(f) \lesssim \epsilon(n, s)$$

where $\epsilon(\theta, a)$ decays with n increasing and rate depending on s, e.g. $n^{-\frac{2s}{2s+d}}$

UniGe | MaLGa

Outline

Data driven modeling paradigm

Statistical learning, bias and variance

From statistics to optimization

A theory crisis?

Enter optimization

$$\min_{\|w\| \leqslant \theta} \widehat{L}(f_w)$$

Enter optimization

$$\min_{\|w\| \leqslant \theta} \widehat{L}(f_w)$$

Gradient methods

$$\widehat{w}_{\theta,t+1} = P_\theta \left(\widehat{w}_{\theta,t} - \gamma_t \nabla \widehat{L}(f_{\widehat{w}_{\theta,t}}) \right)$$

Hard to prove convergence unless the model is linear...

Linear models

$$f_w = \sum_{j=1}^{\infty} w^j \phi_j \quad + \quad \ell \quad \text{convex}$$

$$\Longrightarrow \min_{\|w\| \leqslant \theta} \widehat{L}(f_w) \quad \text{convex problem!}$$

Linear models

$$f_w = \sum_{j=1}^{\infty} w^j \phi_j \quad + \quad \ell \quad \text{convex}$$

$$\Longrightarrow \min_{\|w\| \leqslant \theta} \widehat{L}(f_w) \quad \text{convex problem!}$$

Optimization results ensures that $\widehat{f}_{\theta,t} = f_{\widehat{w}_{\theta,t}}$

$$\widehat{L}(\widehat{f}_{\theta,t}) - \min_{\|w\| \leqslant \theta} \widehat{L}(f) \leqslant \delta_t$$

with δ_t decaying in t.

[Nesterov, '03]

Estimation + optimization error

Lemma

$$L(\widehat{f}_{\theta,t}) - L(f_\theta) \leqslant C_\ell \sup_{\|w\| \leqslant \theta} |\widehat{L}(f_w) - L(f_w)| + \delta_t$$

Estimation + optimization error

Lemma

$$L(\widehat{f}_{\theta,t}) - L(f_\theta) \leqslant C_\ell \sup_{\|w\| \leqslant \theta} |\widehat{L}(f_w) - L(f_w)| + \delta_t$$

Proof.

$$L(\widehat{f}_{\theta,t}) - L(f_\theta) = L(\widehat{f}_{\theta,t}) - \widehat{L}(\widehat{f}_{\theta,t}) + \underbrace{\widehat{L}(\widehat{f}_{\theta,t}) - \widehat{L}(f_\theta)}_{\leqslant \delta_t} + \widehat{L}(\widehat{f}_\theta) - L(f_\theta)$$

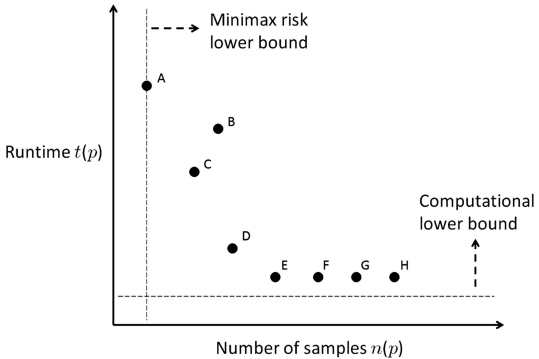
□

The statistical-optimization trade-off

$$L(\widehat{f}_{\theta,t}) - \min L(f) \lesssim \frac{\theta}{\sqrt{n}} + a(\theta, s) + \delta_t$$

The statistical-optimization trade-off

$$L(\widehat{f}_{\theta,t}) - \min L(f) \lesssim \frac{\theta}{\sqrt{n}} + a(\theta, s) + \delta_t$$



[Bousquet, Bottou '07, Chendrasekaran Jordan, '12]

Outline

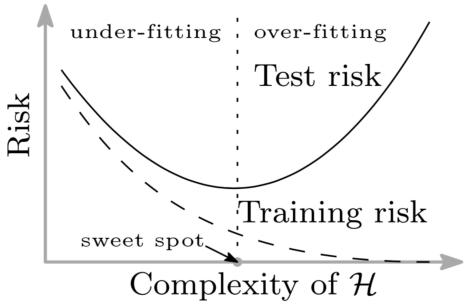
Data driven modeling paradigm

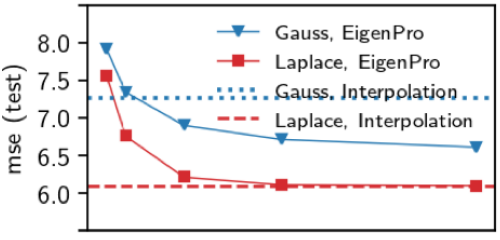
Statistical learning, bias and variance

From statistics to optimization

A theory crisis?

Was it all wrong?





Explicit regularization

$$\min_{\|w\| \leqslant \theta} \widehat{L}(f_w) \qquad \widehat{w}_{\theta,t+1} = P_\theta \left(\widehat{w}_{\theta,t} - \gamma_t \nabla \widehat{L}(f_{\widehat{w}_{\theta,t}}) \right)$$

Explicit regularization

$$\min_{\|w\| \leqslant \theta} \widehat{L}(f_w) \qquad \widehat{w}_{\theta,t+1} = P_\theta \left(\widehat{w}_{\theta,t} - \gamma_t \nabla \widehat{L}(f_{\widehat{w}_{\theta,t}}) \right)$$

Implicit regularization

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \nabla \widehat{L}(f_{\widehat{w}_t})$$

Explicit regularization

$$\min_{\|w\| \leqslant \theta} \widehat{L}(f_w) \qquad \widehat{w}_{\theta,t+1} = P_\theta \left(\widehat{w}_{\theta,t} - \gamma_t \nabla \widehat{L}(f_{\widehat{w}_{\theta,t}}) \right)$$

Implicit regularization

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \nabla \widehat{L}(f_{\widehat{w}_t})$$

Can we characterize $\widehat{f}_t = f_{\widehat{w}_t}$

$$L(\widehat{f}_t) - L(f_*)$$

Inexact optimization with linear models

Linear models: $f_w = \sum_{j=1}^{\infty} w^j \phi_j$, ℓ convex and

$$w_{t+1} = w_t - \gamma_t \nabla L(f_{w_t}),$$

then for $f_t = f_{w_t}$

$$L(f_t) - L(f_*) \leqslant \delta_t.$$

Inexact optimization with linear models

Linear models: $f_w = \sum_{j=1}^{\infty} w^j \phi_j$, ℓ convex and

$$w_{t+1} = w_t - \gamma_t \nabla L(f_{w_t}),$$

then for $f_t = f_{w_t}$

$$L(f_t) - L(f_*) \leqslant \delta_t.$$

Idea: consider

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \left(\nabla L(f_{\widehat{w}_t}) + e_t\right)$$

with

$$e_t = \widehat{\nabla L}(f_{\widehat{w}_{\theta,t}}) - \nabla L(f_{\widehat{w}_{\theta,t}}).$$

[Rockafellar, '76, Salzo, Villa '11, Schmidt, Le Roux, Bach '11]

Excess risk control with inexact gradient

Lemma

$$L(\widehat{f}_t) - L(f_*) \leqslant \delta_t + \sum_{j=1}^{t} \left\langle e_t, \widehat{f}_t - f_* \right\rangle .$$

Excess risk control with inexact gradient

Lemma

$$L(\widehat{f}_t) - L(f_*) \leqslant \delta_t + \sum_{j=1}^{t} \left\langle e_t, \widehat{f}_t - f_* \right\rangle .$$

Need to control:

- gradient error e_t,
- path $(\widehat{f}_j)_j$ around f_*.

Gradient concentration

$$\mathbb{E}\left[\sup_{\|w\|\leqslant\theta}\|\nabla\widehat{L}(f_w)-\nabla L(f_w)\|\right]\lesssim\frac{\theta}{\sqrt{n}}$$

Gradient concentration

$$\mathbb{E}\left[\sup_{\|w\|\leqslant\theta}\|\nabla\widehat{L}(f_w)-\nabla L(f_w)\|\right]\lesssim\frac{\theta}{\sqrt{n}}$$

Path control

For $j\lesssim\sqrt{n}$

$$\|\widehat{f}_t-f_*\|\lesssim\|f_*\|.$$

[Stankewitz, Mücke, R. '21, see also Lin R. '17]

Excess risk control with inexact gradient

Theorem (Stankewitz, Mücke, R. '21)

] For $t \lesssim \sqrt{n}$,

$$\mathbb{E}\left[L(\widehat{f}_t) - L(f_*)\right] \lesssim \frac{1}{\sqrt{n}}$$

Same as explicit regularization: implicit regularization a *new* algorithmic idea[1].

[1] In inverse problem the idea is known since the '50s as iterative regularization …

”Looking for the lost keys under the lamp, because that’s where the light is.”, Yann Lecun

- Can we explain the lack of variance? Learning & interpolation?
- **Are linear model of any practical use?**
- Can linear model explain deep learning?

ML meets large scale computing

Scalable implementations needed $\mapsto$ FALKON

Function `Falkon`$(X \in \mathbb{R}^{n\times d}, y \in \mathbb{R}^n, \lambda, m, t)$**:**

$X_m \leftarrow$ `RamdomSubsample`(X, m);

$T, A \leftarrow$ `Preconditioner`(X_m, λ);

Function `LinOp`(β)**:**

$v \leftarrow A^{-1}\beta$;

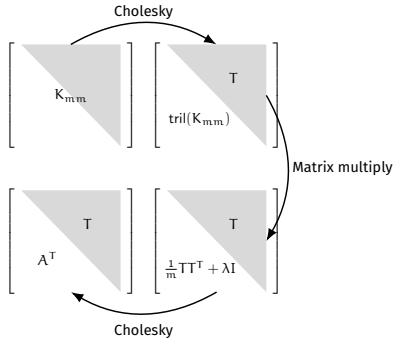
$c \leftarrow k(X_m, X)k(X, X_m)T^{-1}v$;

return $A^{-\top}T^{-\top}c + \lambda n v$;

rhs $\leftarrow A^{-\top}T^{-\top}k(X, X_m)y$;

$\beta \leftarrow$ `ConjugateGradient`$($`LinOp`, rhs, $t)$;

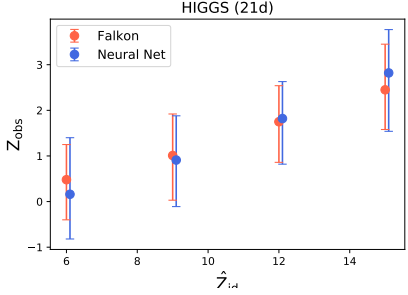
return $T^{-1}A^{-1}\beta$;



[Meanti, Carratino, R., Rudi '20, Meanti, Carratino, De Vito, R. '21]

Efficient linear models in practice: HEP

[Wulzer, D'Agnolo '18]



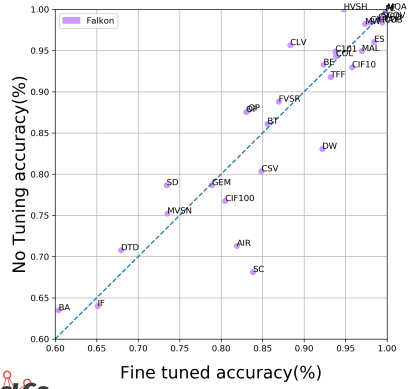
Model	DIMUON	SUSY	HIGGS
Falkon	(53.8 ± 1.9) s	(44.8 ± 1.5) s	(88.7 ± 2.2) s
Neural Net	(4.23 ± 0.73) h	(73.1 ± 10) h	(112 ± 9) h

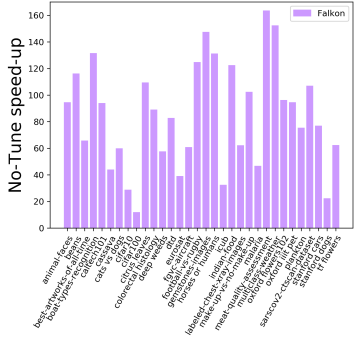
Table 4: Average training times per single run with standard deviations.

[Letizia et al. '21]

Efficient linear models in practice: vision

$$f(x) = \langle w, \Phi(x) \rangle, \quad x \mapsto \underbrace{\Phi_L}_{\text{Efficient linear learning}} \circ \underbrace{\Phi_{L-1} \cdots \circ \Phi_1(x)}_{\text{Pre-trained layers}}$$





[Alfano, Pastore, Odone, R. '21]

- Can we explain the lack of variance? Learning & interpolation.

- Are linear model of any practical use?

- **Can linear model explain deep learning?**

Neural nets as a linear model?

$$f(x) = \sum_{j=1}^{p} \beta^j \sigma(a_j{}^\top x + \alpha_j)$$

Neural nets as a linear model?

$$f(x) = \sum_{j=1}^{p} \beta^j \sigma(a_j{}^\top x + \alpha_j)$$

Key idea: Let $\rho(x, \underbrace{(a, \alpha)}_{\xi}) = \sigma(a_j{}^\top x + \alpha_j)$.

$$f_\mu(x) = \int \rho(x, \xi) d\mu(\xi), \qquad \mu = \sum_{j=1}^{q} \delta_{\xi_j} \beta_j$$

[Bach '14, Montanari et al. 18, ...]

Optimality of neural nets

$$f_\mu(x) = \int \rho(x, \xi) d\mu(\xi) = \langle \rho(x, \cdot), \mu \rangle \,, \qquad \|f\|_{\mathcal{B}} = \|\mu\|_{TV}$$

Optimality of neural nets

$$f_\mu(x) = \int \rho(x, \xi) d\mu(\xi) = \langle \rho(x, \cdot), \mu \rangle, \qquad \|f\|_{\mathcal{B}} = \|\mu\|_{TV}$$

Theorem (Bartolucci, De Vito, R., Vigogna, '21)

The problem

$$\min_{f \in \mathcal{B}} \widehat{L}(f) + \|f\|_{\mathcal{B}}$$

has a solution of the form

$$f(x) = \sum_{i=1}^{u} \beta^j \rho(x, \lambda_i), \quad u \leqslant n.$$

see also [Bredies and Carioni '20, Unser et al. '17, Unser '20, Parhi and Nowak '20].

Wrapping up

- From bias-variance…
- …to statistics-optimization trade-off.
- Statistical learning 2.0?

What's next?

- Physics (equation?) informed ML.
- ML for Inverse problems.
- BIGGER linear models: ML meets scientific computing/HPC.

Multiple post-docs/PhD positions @MaLGa!

@lrntzrsc

malga.unige.it

UniGe | MaLGa