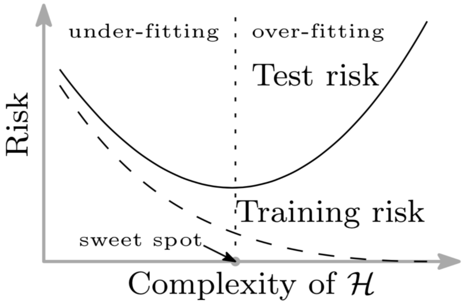
Implicit regularization and acceleration in machine learning

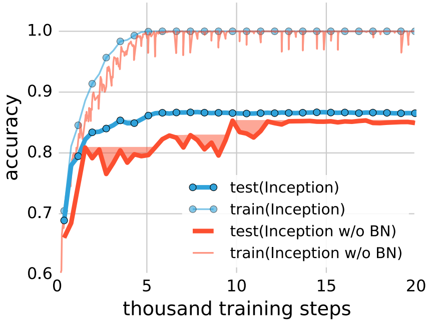
Lorenzo Rosasco
MaLGa- Machine learning Genova Center
Università di Genova
MIT
IIT

Joint work with: N. Pagliana, MaLGa - DIMA - UniGe

Mathematics of Deep Learning - TU Delft

There seems to be a puzzle





Outline

Optimization for machine learning

Part I: Learning theory of (accelerated) optimization

Part II: More learning theory and some science of (accelerated) optimization

- Refined results: easy problems
- Refined results: hard problems

Optimization for machine learning

Training error

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \lambda \|w\|^2$$

Gradient methods

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(f_{\widehat{w}_t}(x_i), y_i) - 2\gamma_t \lambda \widehat{w}_t$$

Optimization for machine learning

Training error

$$\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \lambda\|w\|^2$$

Gradient methods

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{n}\sum_{i=1}^{n} \nabla\ell(f_{\widehat{w}_t}(x_i), y_i) - 2\gamma_t\lambda\widehat{w}_t$$

$$\lim_{t\to\infty} \frac{1}{n}\sum_{i=1}^{n} \ell(f_{\widehat{w}_t}(x_i), y_i) + \lambda\|\widehat{w}_t\|^2 = \min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \lambda\|w\|^2$$

$\Longrightarrow$ **Go faster! ...but where?**

Statistical machine learning

$$\frac{1}{n}\sum_{i=1}^{n}\ell(f_w(x_i), y_i) \approx \mathbb{E}_{x,y}[\ell(f_w(x), y)]$$

Test error

$$\mathbb{E}_{x,y}[\ell(f_{\widehat{w}_t}{}^\top(x), y)]$$

UniGe | MaLGa

Error measures

Generalization error

$$\frac{1}{n}\sum_{i=1}^{n} \ell(f_{\widehat{w}_t}(x_i), y_i) - \mathbb{E}_{x,y}[\ell(f_{\widehat{w}_t}(x), y)]$$

Excess risk

$$\mathbb{E}_{x,y}[\ell(f_{\widehat{w}_t}(x), y)] - \min_{w \in \mathbb{R}^p} \mathbb{E}_{x,y}[\ell(f_w(x), y)]$$

Outline

Optimization for machine learning

Part I: Learning theory of (accelerated) optimization

Part II: More learning theory and some science of (accelerated) optimization

- Refined results: easy problems
- Refined results: hard problems

Least squares learning

Solve

$$\min_{w \in \mathbb{R}^p} \mathbb{E}_{x,y}[(w^\top \Phi(x) - y)^2]$$

where $\Phi(x) \in \mathbb{R}^p$ and p can be infinite.

Gradient descent[1]

$$\widehat{w}_{t+1} = \widehat{w}_t - \alpha \nabla \widehat{L}(\widehat{w}_t), \qquad \nabla \widehat{L}(w) = \frac{2}{n} \sum_{i=1}^{n} \Phi(x_i)(w^\top \Phi(x_i) - y_i)$$

with

$$\alpha = \frac{1}{\sup_x \|\Phi(x)\|^2}.$$

[1] $(x_1, y_1), \ldots, (x_n, y_n)$ iid.

Accelerated iterations

Heavy-ball

$$\widehat{w}_{t+1} = \widehat{w}_t - \alpha_t \nabla \widehat{L}(\widehat{w}_t) + \beta_t (\widehat{w}_t - \widehat{w}_{t-1}).$$

[2] Reduces to Chebyshev iterative method for $\nu = 1/2$.

Accelerated iterations

Heavy-ball

$$\widehat{w}_{t+1} = \widehat{w}_t - \alpha_t \nabla \widehat{L}(\widehat{w}_t) + \beta_t (\widehat{w}_t - \widehat{w}_{t-1}).$$

In particular[2] for $\nu > 0$

$$\alpha_t = \frac{1}{\sup_x \|\Phi(x)\|^2} \frac{4(2t+2\nu-1)(t+\nu-1)}{(t+2\nu-1)(2t+4\nu-1)}, \qquad \beta_t = \frac{(t-1)(2t-3)(2t+2\nu-1)}{(t+2\nu-1)(2t+4\nu-1)(2t+2\nu-3)}.$$

UniGe | MaLGa

[2] Reduces to Chebyshev iterative method for $\nu = 1/2$.

Accelerated iterations

Heavy-ball

$$\widehat{w}_{t+1} = \widehat{w}_t - \alpha_t \nabla \widehat{L}(\widehat{w}_t) + \beta_t(\widehat{w}_t - \widehat{w}_{t-1}).$$

In particular[2] for $\nu > 0$

$$\alpha_t = \frac{1}{\sup_x \|\Phi(x)\|^2} \frac{4(2t+2\nu-1)(t+\nu-1)}{(t+2\nu-1)(2t+4\nu-1)}, \qquad \beta_t = \frac{(t-1)(2t-3)(2t+2\nu-1)}{(t+2\nu-1)(2t+4\nu-1)(2t+2\nu-3)}.$$

Nesterov's acceleration

$$\widehat{w}_{t+1} = \widehat{v}_t - \alpha \nabla \widehat{L}(\widehat{v}_t), \qquad \widehat{v}_t = \widehat{w}_t + \beta_t (\widehat{w}_t - \widehat{w}_{t-1}).$$

In particular for $\beta > 1$

$$\alpha = \frac{1}{\sup_x \|\Phi(x)\|^2}, \qquad \beta_t = \frac{t-1}{t+\beta}.$$

[2] Reduces to Chebyshev iterative method for $\nu = 1/2$.

Basic result

Let

$$L(w) = \mathbb{E}_{x,y}[(w^\top x - y)^2], \qquad L(w_*) = \min_{w \in \mathbb{R}^p} L(w)$$

Theorem

Assume $\|\Phi(x)\|, |y| \leqslant 1$ a.s.. Then w.h.p.

$$L(\widehat{w}_t) - L(w_*) \lesssim \frac{1}{t} + \frac{t}{n}$$

for GD, whereas

$$L(\widehat{w}_t) - L(w_*) \lesssim \frac{1}{t^2} + \frac{t^2}{n}$$

for Heavy-ball and Nesterov acc.

Basic result (cont.)

Corollary

For GD, if $t = \sqrt{n}$*,*

$$L(\widehat{w}_t) - L(w_*) \lesssim \frac{1}{\sqrt{n}}.$$

The same bound hold for for Heavy-ball and Nesterov acc. for $t = n^{1/4}$.

Numerical illustration

Parameters of the plot in the left: space size $N = 10^4$, training points $n = 10^2$, $\gamma = 1$, noise $\sigma = 0.5$, step-size $\alpha \ll 0.9/\max(eigs(\widehat{K})) \leqslant \frac{1}{\sup_x \|\Phi(x)\|^2}$.

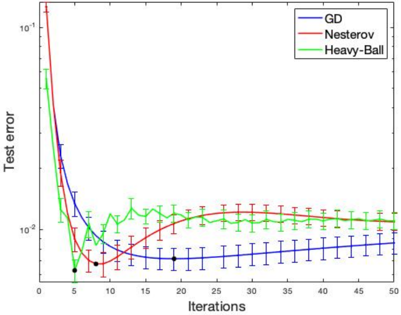


Figure: Simulated data (ill-conditioned LS)

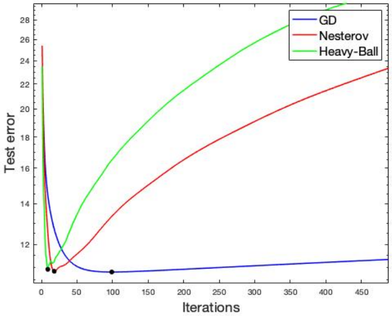


Figure: Pumadyn8nh dataset ($n = 8192$, $d = 7$), Gaussian kernel width1.2.

Remarks

- ~~Early stop after $\sqrt{n}$ iteration!~~ Iterations control complexity/stability.
- Acceleration can suffer from instability.
- Iterates converge to minimal norm minimizer (implicit bias).
- Training error/generalization play no role.
- Proof based on spectral filtering/calculus (Engl et al. '96, Neubauer '16)

Remarks

- ~~Early stop after $\sqrt{n}$ iteration!~~ Iterations control complexity/stability.
- Acceleration can suffer from instability.
- Iterates converge to minimal norm minimizer (implicit bias).
- Training error/generalization play no role.
- Proof based on spectral filtering/calculus (Engl et al. ’96, Neubauer ’16)

We can see other behaviors in practice: explanation?

Outline

Optimization for machine learning

Part I: Learning theory of (accelerated) optimization

Part II: More learning theory and some science of (accelerated) optimization

- Refined results: easy problems
- Refined results: hard problems

Do we like assumptions or not?

- ”Simple and Almost Assumption-Free Out-of-Sample Bound for ...”
- ”...a more ambitious open problem (to find good bounds) is to find the correct characterization of “easiness” for real-world problem...”

UniGe | MaLGa

Do we like assumptions or not?

- ”Simple and Almost Assumption-Free Out-of-Sample Bound for ...”
- ”...a more ambitious open problem (to find good bounds) is to find the correct characterization of “easiness” for real-world problem...”

We can see other behaviors in practice: explanation?

Refined assumption: easy problems

$$\Sigma = \mathbb{E}_x[\Phi(x)\Phi(x)^\top] \qquad h = \mathbb{E}_{x,y}[\Phi(x)y]$$

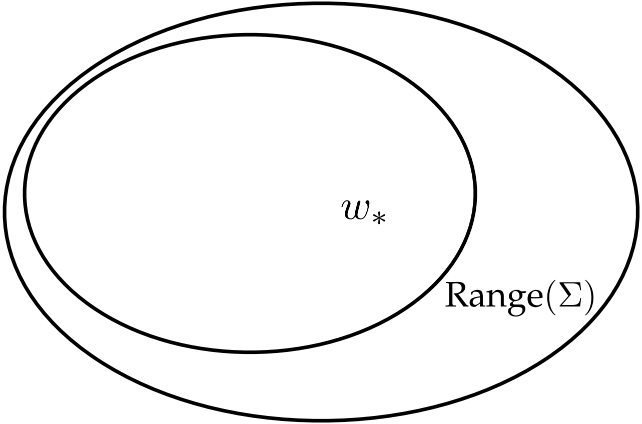
Optimality condition

$$L(w_*) = \min_{w \in \mathbb{R}^p} \mathbb{E}_{x,y}[(w^\top \Phi(x) - y)^2] \quad \Leftrightarrow \quad \Sigma w_* = h.$$

Error/source condition

$$w_* \in \mathrm{Range}(\Sigma^s), \qquad s \in [0, \infty)$$

Easy problems illustrated



Refined results

Theorem

Under the error/source condition, assume $\|\Phi(x)\|, |y| \leqslant 1$ *a.s.. Then w.h.p.*

$$L(\widehat{w}_t) - L(w_*) \lesssim \frac{1}{t^{2s+1}} + \frac{t}{n}$$

with $s \in [0, \infty)$ *for GD, whereas*

$$L(\widehat{w}_t) - L(w_*) \lesssim \frac{1}{t^{2(2s+1)}} + \frac{t^2}{n}$$

with $s \in [0, \nu)$ *for Heavy-ball and with* $s = 0$ *for Nesterov acc.*

Refined results (cont.)

Corollary

For GD with $s \in [0, \infty)$, *choosing* $t = n^{\frac{1}{2s+2}}$,

$$L(\widehat{w}_t) - L(w_*) \lesssim \frac{1}{n^{\frac{2s+1}{2s+2}}}.$$

The same bound hold for Heavy-ball with $s \in [0.\nu)$ *and for Nesterov acc. with* $s = 0$ *choosing* $t = \sqrt{n^{\frac{1}{(2s+2)}}}$.

Acceleration can suffer from slow rates for easy problems.

Numerical illustration

Parameters: space size $N = 10^4$, training points $n = 10^2$, $\gamma = 1$, noise $\sigma = 0.2$, step-size $\alpha = 0.9/\max(eigs(\widehat{K})) \leqslant \frac{1}{\sup_x \|\Phi(x)\|^2}$.

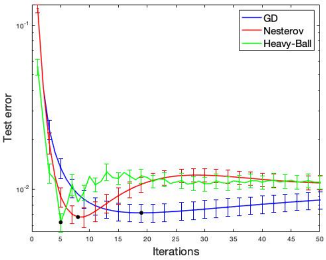


Figure: $s = 0$

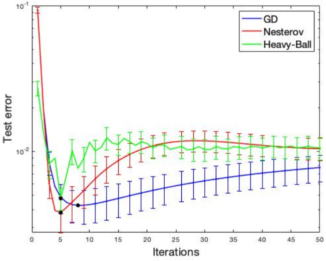


Figure: $s = 3/2$

UniGe | MaLGa

Numerical illustration

Parameters: space size $N = 10^4$, training points $n = 10^2$, $\gamma = 1$, noise $\sigma = 0.5$, step-size $\alpha = 0.9/\max(eigs(\widehat{K})) \leqslant \frac{1}{\sup_x \|\Phi(x)\|^2}$.

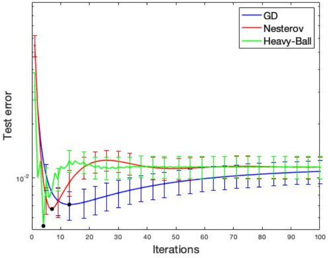


Figure: $s = 0$

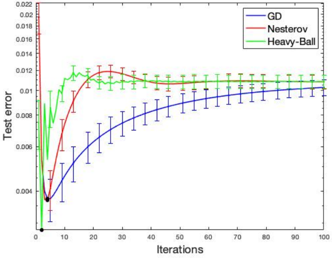


Figure: $s = 50$

So far

- Large function class/simple target function: instability and slow rate?

  Gradient descent might catch up.

- What about small function class/complex target function?

Refined assumption: hard problems

Let

$$\overline{\Sigma} f(x) = \mathbb{E}_x[\Phi(x) f(x)]$$

General source condition

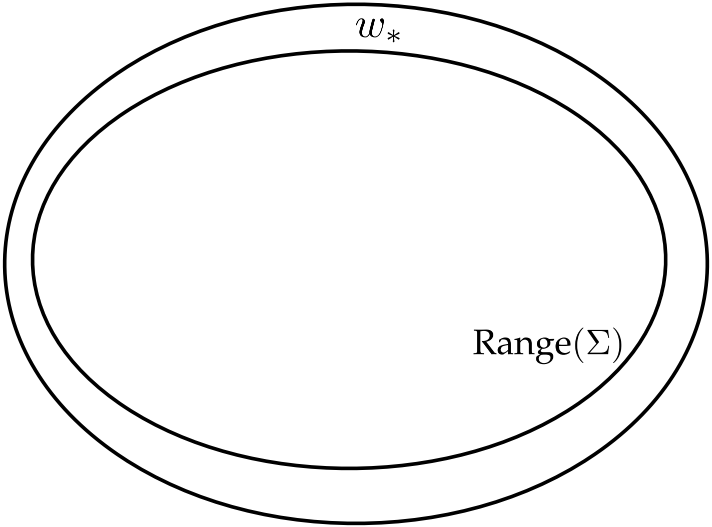
$$\mathbb{E}[y|x] \in \text{Range}(\log(\overline{\Sigma})).$$

Eigendecay

$$\sigma_j(\Sigma) \sim e^{-j}.$$

Example: Learn a smooth (Sobolev) function with a Gaussian kernel (fixed width!).

Hard problems illustrated



Refined results

Theorem

Under the error/source condition, assume $\|\Phi(x)\|, |y| \leqslant 1$ *a.s.. Then w.h.p.*

$$L(\widehat{w}_t) - L(w_*) \lesssim \frac{1}{\log(t)} + \frac{\log(t)}{n} + \frac{t}{n^2}$$

with for GD, whereas for

$$L(\widehat{w}_t) - L(w_*) \lesssim \frac{1}{2\log(t)} + \frac{2\log(t)}{n} + \frac{t^2}{n^2}$$

for Heavy-ball and for Nesterov acc.

Refined results (cont.)

Corollary

For GD choosing $t \sim n^{\alpha}$*,* $\alpha < 2$

$$L(\widehat{w}_t) - L(w_*) \lesssim \frac{1}{\log(n)}.$$

The same bound hold for Heavy-ball and for Nesterov acc. with $t \sim \sqrt{n^{\alpha}}$*,* $\alpha < 2$*.*

Numerical illustration

Parameters: space size $N = 10^4$, training points $n = 10^2$, $\gamma = 1$, source condition logarithmic, noise $\sigma = 0.2$, step-size $\alpha = 0.9/\max(eigs(\widehat{K})) \leqslant \frac{1}{\sup_x \|\Phi(x)\|^2}$.

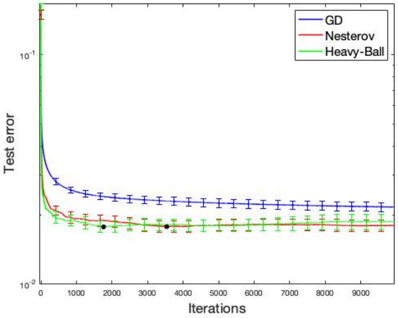


Figure: Simulation of the test error in the case $\sigma_i \approx e^{-\gamma i}$

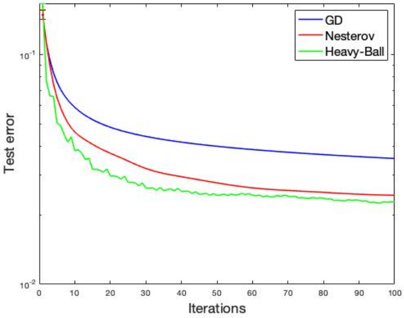


Figure: Simulation of the test error in the case $\sigma_i \approx e^{-\gamma i}$ (zoom)

ML Science

The behavior of an algorithms depending on modeling assumptions.

Which assumptions are good depends on data.

Looking at different assumptions allows to explaning different empirical behaviors.

Wrapping up

- Optimization for machine leads to new algorithms: implicit regularization.
- Different behaviors depending on easy/hard learning problems.
- TBD: high/low dimension and SNR, classification; nonlinear parameterization...

$$L(\widehat{w}_t) - L(w_*) \lesssim \frac{1}{\log(t)} + \frac{\log(t)}{n} + \frac{\overset{\text{const.}}{\cancel{t}}}{n^2} \ ?$$

UniGe | MaLGa New machine learning center in Genova: PhD positions available!

Outline

Spectral filtering

Spectral filtering

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^\top, \qquad \widehat{h} = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i)y_i$$

GD Filter

$$\widehat{w}_{t+1} = \widehat{w}_t - \alpha\nabla\widehat{L}(\widehat{w}_t) = \alpha\sum_{j=0}^{t}(I - \alpha\widehat{\Sigma})^j\widehat{h}$$

For t large,

$$g_t(\widehat{\Sigma}) = \alpha\sum_{j=0}^{t}(I - \alpha\widehat{\Sigma})^j \approx \widehat{\Sigma}^{-1}$$

$$g_t(\widehat{\Sigma}) = \alpha\sum_{j=0}^{t}(I - \alpha\widehat{\Sigma})^j \approx \widehat{\Sigma}^{-1}$$

Spectral filters

Definition

$\{g_\lambda\}_{\lambda\in(0,1]}$ is a *spectral filtering function* if there exists $E, F_0, q, (F_s)_{s=0}^q < \infty$ s.t., for any $\lambda \in (0,1]$

(i)

$$\sup_{\sigma\in(0,\kappa^2]} |g_\lambda(\sigma)| \leqslant \frac{E}{\lambda} .$$

(ii) Let $r_\lambda(\sigma) = 1 - \sigma\, g_\lambda(\sigma)$, for $s \in [0, q)$

$$\sup_{\sigma\in(0,\kappa^2]} |r_\lambda(\sigma)\sigma^s| \leqslant F_s \lambda^s .$$

The parameter q is called qualification.

Nicoló Pagliana, Lorenzo Rosasco, Implicit Regularization of Accelerated Methods in Hilbert Spaces, arxiv.