Inverse Problems in Machine Learning

Algorithmic Regularization for Machine Learning

Luigi Carratino
University of Genoa

joint work with Alessandro Rudi (INRIA), Lorenzo Rosasco (UniGe, MIT, IIT)

Mar, 1st 2019 – SIAM CSE 2019

Learning algorithms design

1. Statistical estimation: minimization of an empirical objective

2. Optimization

What is the effect of optimization on the statistical properties?

Statistical Learning

Let $(x, y) \sim \rho$, $x \in \mathbb{R}^d$, $y \in \mathbb{R}$, $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ hilbert space, $\phi : \mathbb{R}^d \to \mathcal{H}$,

The problem

Learn a non-linear function $f : \mathbb{R}^d \to \mathbb{R}$ e.g. $f(x) = \langle w, \phi(x) \rangle$, solving

$$\min_{w \in \mathcal{H}} \mathcal{E}(w), \qquad \mathcal{E}(w) = \int (y - \langle w, \phi(x) \rangle)^2 d\rho(x, y)$$

with ρ **unknown**, given a set of samples $(x_i, y_i)_{i=1}^n \sim \rho^n$.

Statistics

$$\widehat{w}_\lambda = \operatorname*{argmin}_{w \in \mathcal{H}} \widehat{\mathcal{E}}(w), \qquad \widehat{\mathcal{E}}(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, \phi(x_i) \rangle)^2 + \lambda \|w\|_{\mathcal{H}}^2$$

Statistics

$$\widehat{w}_\lambda = \operatorname*{argmin}_{w \in \mathcal{H}} \widehat{\mathcal{E}}(w), \qquad \widehat{\mathcal{E}}(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, \phi(x_i) \rangle)^2 + \lambda \|w\|_{\mathcal{H}}^2$$

Theorem (Caponnetto, De Vito '05)

For $\|x\|, |y| \leq 1$ a.s. and $\lambda = \frac{1}{\sqrt{n}}$

$$\mathcal{E}(\widehat{w}_{\lambda_n}) - \inf_{w \in \mathcal{H}} \mathcal{E}(w) \lesssim \frac{1}{\sqrt{n}}$$

Optimization

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma \nabla \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w_t, \phi(x_i) \rangle)^2 + \lambda \|w_t\|^2 \right)$$

Theorem

If $\gamma \leq 1$, *then*

$$\widehat{\mathcal{E}}(w_t) - \widehat{\mathcal{E}}(w_\lambda) \lesssim e^{-t\lambda}$$

Computational tricks = (implicit) regularization?

- **iterations**
- acceleration
- **stochastic gradients**
- **step-size**
- **mini-batch**
- averaging
- **sketching**
- subsampling
- preconditioning
- ...

Random features

Let $f(x)$ be

$$f(x) = \langle w, \phi_M(x) \rangle$$

where $\phi_M : \mathbb{R}^d \to \mathbb{R}^M$

$$\phi_M(x) := \left(\underbrace{\sigma(\langle x, s_1 \rangle)}_{\text{random feature}}, \ldots, \sigma(\langle x, s_M \rangle) \right)$$

- $s_1, \ldots, s_M \in \mathbb{R}^d$ i.i.d random vectors
- $\sigma : \mathbb{R} \to \mathbb{R}$ nonlinear function (e.g. $\sigma(a) = cos(a)$, $\sigma(a) = |a|_+, \ldots$)

[Rahimi, Recht '06'08'08]

Random features

Note:

- Neural netowrk with random weights

$$f(x) = \langle w, \phi_M(x) \rangle = \sum_{j=1}^{M} w^j \sigma(\langle s_j, x \rangle)$$

- As $M \to \infty$, for p.d. kernel $K : X \times X \to \mathbb{R}$

$$\langle \phi_M(x), \phi_M(x') \rangle \approx \langle \phi(x), \phi(x') \rangle = K(x, x')$$

SGD with Random Features

For $t = 1, \ldots, T$

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \nabla \left(\left(y_t - \langle \widehat{w}_t, \phi_M(x_t) \rangle \right)^2 \right)$$

with $(x_1, y_1), \ldots, (x_t, y_t)$ sampled uniformly at random from $(x_i, y_i)_{i=1}^n$.

SGD-RF with mini-batching

For $t = 1, \dots, T$

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} \nabla \left(\left(y_{j_i} - \langle \widehat{w}_t, \phi_M(x_{j_i}) \rangle \right)^2 \right)$$

with $j_1, \dots, j_{bT}$ sampling strategy.

SGD-RF with mini-batching

For $t = 1, \ldots, T$

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} \nabla \left(\left(y_{j_i} - \langle \widehat{w}_t, \phi_M(x_{j_i}) \rangle \right)^2 \right)$$

with $J = j_1, \ldots, j_{bT}$ sampling strategy.

Free parameters:

- Step-size γ_t
- Mini-batch size b
- Number of random features M
- Number of iterations T

Computational complexity:

- Time: $O(MbT)$
- Space: $O(M)$

Previous results

- One pass SGD: from Robbins-Munro ’50’s... Dieuleveut, Bach ’15...
- Multipass SGD: Hardt Recht Singer ’16, Rosasco et al. ’16
- Sketching for Tikhonov regularization: Rudi, Rosasco ’17.
- Multipass SGD+Sketching: This work!

SGD with Random Features: Statistics

Theorem (C., Rudi, Rosasco '18)

For $|\sigma(\langle x, s\rangle)|, |y| \leq 1$, $t > 1$ with probability $1 - \delta$

$$\mathbb{E}_J \mathcal{E}(\widehat{w}_{t+1}) - \inf_{w \in \mathcal{H}} \mathcal{E}(w) \lesssim \frac{\gamma}{b} + \left(\frac{\gamma t}{M} + 1\right) \frac{\gamma t \log \frac{1}{\delta}}{n} + \frac{\log \frac{1}{\delta}}{M} + \frac{1}{\gamma t}.$$

SGD with Random Features: Statistics

Theorem (C., Rudi, Rosasco '18)

If

1. $b = 1$, $\gamma_t \simeq \frac{1}{\sqrt{n}}$, and $T = n$ iterations (1 pass over the data);

and

$$M = \sqrt{n}$$

then with high probability

$$\mathbb{E}_J \mathcal{E}(\widehat{w}_T) - \inf_{w \in \mathcal{H}} \mathcal{E}(w) \lesssim \frac{1}{\sqrt{n}}$$

SGD with Random Features: Statistics

Theorem (C., Rudi, Rosasco '18)

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);

and

$$M = \sqrt{n}$$

then with high probability

$$\mathbb{E}_J \mathcal{E}(\widehat{w}_T) - \inf_{w \in \mathcal{H}} \mathcal{E}(w) \lesssim \frac{1}{\sqrt{n}}$$

SGD with Random Features: Statistics

Theorem (C., Rudi, Rosasco '18)

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);

and

$$M = \sqrt{n}$$

then with high probability

$$\mathbb{E}_J \mathcal{E}(\widehat{w}_T) - \inf_{w \in \mathcal{H}} \mathcal{E}(w) \lesssim \frac{1}{\sqrt{n}}$$

Computational requirements

For $t = 1, \ldots, T$

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} \nabla \left(\left(y_{j_i} - \langle \widehat{w}_t, \phi_M(x_{j_i}) \rangle \right)^2 \right)$$

Complexity:

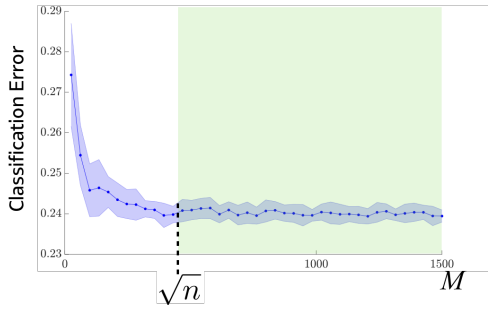
- Time: $O(MbT)$
- Space: $O(M)$

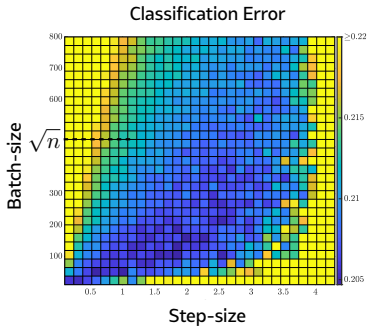
Complexity for $O(1/\sqrt{n})$ rate:

- Time: $O(n\sqrt{n})$
- Space: $O(\sqrt{n})$

Empirical results

SUSY dataset, $n = 6 \times 10^6$





- Same accuracy for $M \geq \sqrt{n}$
- $b = \sqrt{n}$ is the ”magic” MB-size

Summing up

- number of passes, step-size mini-batch size and sketching dimension....
  all control the test error!
- They introduces an implicit bias hence regularize the solution

- + Fast rates
- + Decreasing Stepsize

Looking ahead: apply/extend these ideas

- Beyond least squares
- Parallelization
- Non convex problems