Algorithmic Regularization for Machine Learning

Luigi Carratino
University of Genoa

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

Oct, 4th 2018 – Poggio Lab

Learning algorithms design

In theory: ERM+Optimization

In practice:

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

What is the effect on the test error?

Statistical Learning

Let $(x, y) \sim \rho$, $x \in \mathbb{R}^d$, $y \in \mathbb{R}$

The problem

Solve

$$\min_f \mathcal{E}(f), \qquad \mathcal{E}(f) = \int (y - f(x))^2 d\rho(x, y)$$

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

Statistics

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

Optimization

$$\widehat{f}_{t+1} = \widehat{f}_t - \gamma \nabla \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - f_t(x_i))^2 + \lambda \|f_t\|^2 \right)$$

Theorem

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

$$\widehat{\mathcal{E}}(f_t) - \widehat{\mathcal{E}}(f_\lambda) \lesssim e^{-t\lambda}$$

Computational tricks = (implicit) regularization?

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

Random features

Let $f \in \mathcal{H}$ be

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

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

$$\phi_M(x) := (\sigma(\langle x, s_1 \rangle), \ldots, \sigma(\langle x, s_M \rangle))$$

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

Random features

Note:

- Neural netowrk with random weights

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

- Kernel as $M \to \infty$

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

SGD with Random Features

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 with Random Features

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

Free parameters:

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

Previous results

- One pass SGD: from Robbins-Munro ’50’s... Bach et al. ’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)

Under basic assumptions, with high probability

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

SGD with Random Features: Statistics

Theorem (C., Rudi, Rosasco '18)

If

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

and

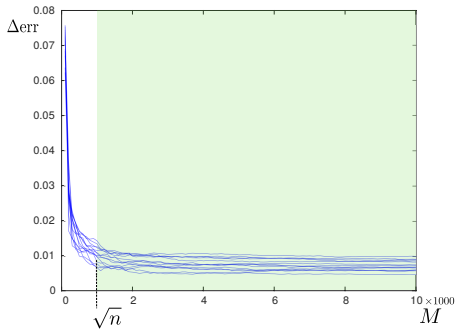
$$M = \sqrt{n}$$

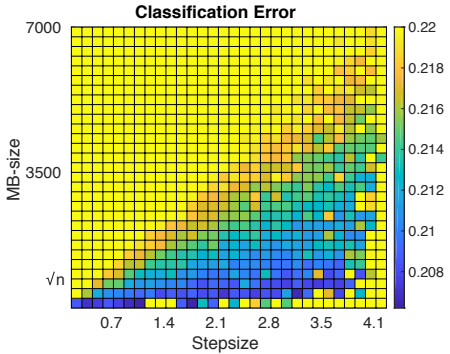
then with high probability

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

Empirical results

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





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

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: $O(M)$
- Eval. of $\phi_M(x)$: $O(M)$

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: $O(M)$
- Eval. of $\phi_M(x)$: $O(M)$

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

- Time: $O(n\sqrt{n})$
- Space: $O(\sqrt{n})$
- Eval. of $\phi_M(x)$: $O(\sqrt{n})$

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

Looking ahead: apply/extend these ideas

- Beyond least squares
- Parallelization
- Non convex problems