Least Squares Learning

Approximately solve

\[ \min_{w} \mathcal{L}(w), \quad \mathcal{L}(w) = \frac{1}{2} \mathbb{E}[(Y - (w, X))^2]. \]

\( \mathcal{H} \): real separable Hilbert space

Define

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

Optimal solution \( w^* \) satisfies normal equation:

\[ w^* = \Sigma h. \]

Mini-Batch SGD Recursion

Let \( t = 0, \ldots, T, w_0 = 0 \) and

\[ w_{t+1} = w_t - \frac{\gamma}{b} \sum_{i=b(t-1)+1}^{b(t)} ((w_i, x_i) - y_i) x_i, \]

\( \{j_1, \ldots, j_{bT} \} \sim \text{i.i.d Unif}[n]. \)

Tail AveSGD: Tail-length \( L = 1, \ldots, T \)

\[ w_t := \frac{1}{T} \sum_{t=T-L+1}^{T} w_t \]

Unif AveSGD: \( L = T \)

Why Tail-Averaging? (Part I)

- too small step sizes \( \rightarrow \) slow convergence
- larger step sizes \( \rightarrow \) improved convergence + noisy trajectories
- Unif AveSGD: allows large/constant step sizes since it reduces the variance of SGD
- Tail AveSGD: sufficiently “long” tail preserve this benefit

Assumption I: Regularity

For some \( r \geq 0 \) we assume \( w \in \text{Run}(\Sigma^r) \).

Note: \( \text{Run}(\Sigma^2) = \mathcal{H} \) and \( \text{Run}(\Sigma^3) \subset \mathcal{H} \)

Main Theorem: Excess Risk of Tail AveSGD

Define effective dimension

\[ N(1/\gamma L) := \text{Tr} \{ \Sigma + 1/(\gamma L)^{1/4} \Sigma \}. \]

Let \( 1 \leq L \leq T \). Assume \( \gamma L < 1/4 \). Then

\[ \mathbb{E}[\mathcal{L}(w_t) - \mathcal{L}(w^*)] \leq \text{Approx}_N(\Sigma, \nu) + \frac{N(1/\gamma L)}{n} + \gamma \text{Tr} \{\Sigma, \nu\} \]

for \( n \) sufficiently large.

Saturation

Let \( \text{Approx}_N(\Sigma, \nu) \) denote the Approximation Error.

Unif AveSGD: \( \text{Approx}_N(\Sigma, \nu) \approx (1/\gamma)^{2m} \]

Tail AveSGD: \( \text{Approx}_N(\Sigma, \nu) \approx (1/\gamma)^{2L} \]

Why Tail-Averaging? (Part II)

- too small step sizes \( \rightarrow \) slow convergence
- larger step sizes \( \rightarrow \) improved convergence + noisy trajectories
- Unif AveSGD: allows large/constant step sizes since it reduces the variance of SGD
- Tail AveSGD: sufficiently “long” tail preserve this benefit

Assumption II: Capacity

For some \( \nu \in (0, 1) \) we assume \( N(1/\gamma L) \lesssim (\gamma L)^{\nu} \).

Corollary: Learning Rate

The excess risk of the (tail)-averaged SGD iterate satisfies

\[ \mathbb{E}[\mathcal{L}(w_t) - \mathcal{L}(w^*)] \lesssim n^{-1/2} \]

for each of the following choices:

- one pass: \( b_n \approx 1, L_n \approx n, \gamma_n \approx n^{-\frac{1}{2}} \)
- one pass: \( b_n \approx n^{-\frac{1}{2}}, L_n \approx n^{-\frac{1}{2}}, \gamma_n \approx 1 \)
- \( \Omega(n^{-\frac{1}{2}}) \) passes: \( b_n \approx n, L_n \approx n^{-\frac{1}{2}}, \gamma_n \approx 1 \)

Experiment: Saturation

Excess risk as a function of regularity \( r \) with uniform and tail averaging.

Unif AveSGD: starts to lag behind its tail-averaged counterpart for larger values of \( r \) exceeding 1/2, flattening out.

Tail AveSGD: continues to improve for large values of \( r \), confirming that this algorithm can indeed massively benefit from favorable structural properties of the data.

Experiment: Single Pass Performance

Single pass performance as a function of the stepsize \( \gamma \) and the minibatch-size \( b \).

Performance: remains largely constant as \( \gamma \cdot b \) remains constant for both algorithms, until a critical threshold stepsize is reached.

Tail AveSGD: permits the use of larger minibatch sizes, allowing for more efficient parallelization.

Acknowledgments: N.M. is supported by the German Research Foundation, DFG Grant STG 1074/4-1. L.B. acknowledges the financial support of the AFOSR projects FA9550-17-1-0390 and BAA-APRL-APOSR-2016-0007 (European Office of Aerospace Research and Development), and the EU H2020-MSCA-RISE project NoMADS - DLV-777826.