Beating SGD Saturation with Tail-Averaging and Minibatching

Nicole Mücke[1] Gergely Neu[2] Lorenzo Rosasco[3]

[1] Universität Stuttgart, [2]Universitat Pompeu Fabra, Barcelona, [3]Universita' degli Studi di Genova, IIT, MIT

Least Squares Learning

Approximately solve

$$\min_{w\in\mathcal{H}} \mathcal{L}(w), \qquad \mathcal{L}(w) = \frac{1}{2}\mathbb{E}[(Y-\langle w, X\rangle)^2] .$$

$\mathcal{H}$: real separable Hilbert space

Define

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

Optimal solution w_* satisfies normal equation:

$$\Sigma w_* = h .$$

Mini-Batch SGD Recursion

Let $t = 0, ..., T$, $w_0 = 0$ and

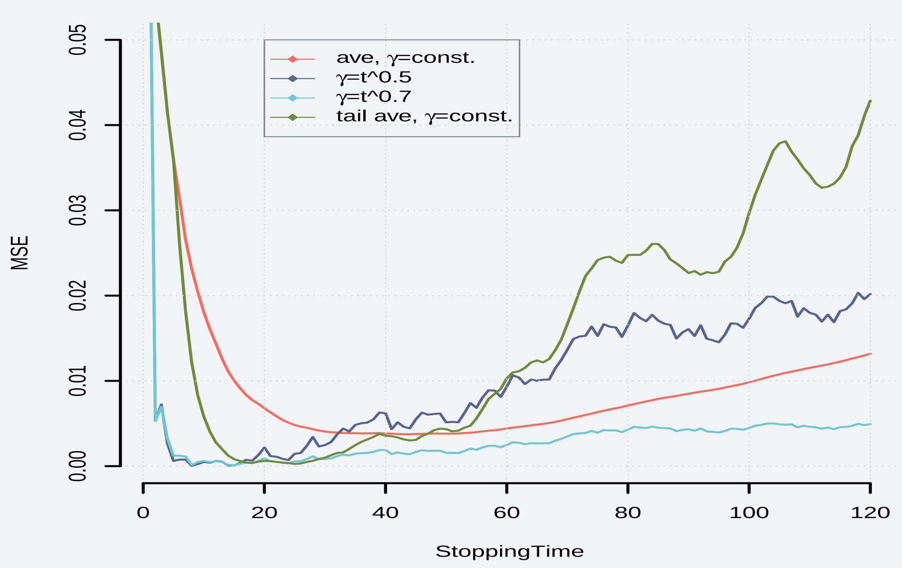
$$w_{t+1} = w_t - \frac{\gamma_t}{b}\sum_{i=b(t-1)+1}^{bt} (\langle w_t, x_{j_i}\rangle - y_{j_i})x_{j_i} ,$$

$j_1, ..., j_{bT} \sim$ i.i.d Unif[n].

Tail AveSGD: Tail-length $L = 1, ..., T$ $\quad \bar{w}_L := \frac{1}{L}\sum_{t=T-L+1}^{T} w_t$

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 $w_* \in \mathrm{Ran}(\Sigma^r)$.

Note: $\mathrm{Ran}(\Sigma^0) = \mathcal{H}$ and $\mathrm{Ran}(\Sigma^r) \subseteq \mathcal{H}$

Main Theorem: Excess Risk of Tail AveSGD

Define effective dimension

$$\mathcal{N}(1/\gamma L) := \mathrm{Trace}[(\Sigma + 1/(\gamma L))^{-1}\Sigma] .$$

Let $1 \leq L \leq T$. Assume $\gamma\kappa^2 < 1/4$. Then

$$\mathbb{E}[\mathcal{L}(\bar{w}_L) - \mathcal{L}(w_*)] \lesssim \mathrm{Approx}_L(\Sigma, w_*) + \frac{\mathcal{N}(1/\gamma L)}{n} + \frac{\gamma\,\mathrm{Trace}[\Sigma^\nu]}{b(\gamma L)^{1-\nu}}$$

for n sufficiently large.

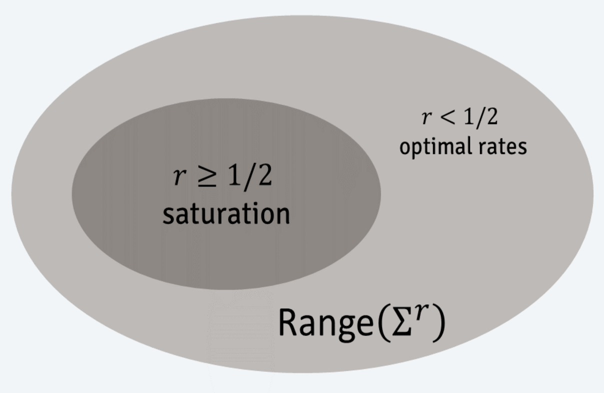
Saturation

Let $\mathrm{Approx}_L(\Sigma, w_*)$ denote the Approximation Error.

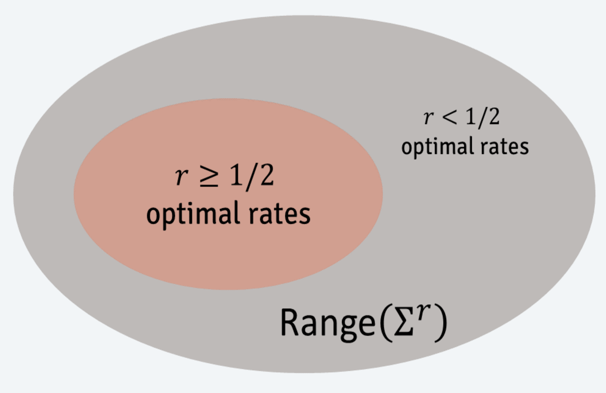
Unif AveSGD: $\mathrm{Approx}_T(\Sigma, w_*) \approx (1/\gamma T)^{2\min(r,1/2)+1}$

Tail AveSGD: $\mathrm{Approx}_L(\Sigma, w_*) \approx (1/\gamma L)^{2r+1}$

Why Tail-Averaging ? (Part II)



Unif AveSGD



Tail AveSGD

Assumption II: Capacity

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

Corollary: Learning Rate

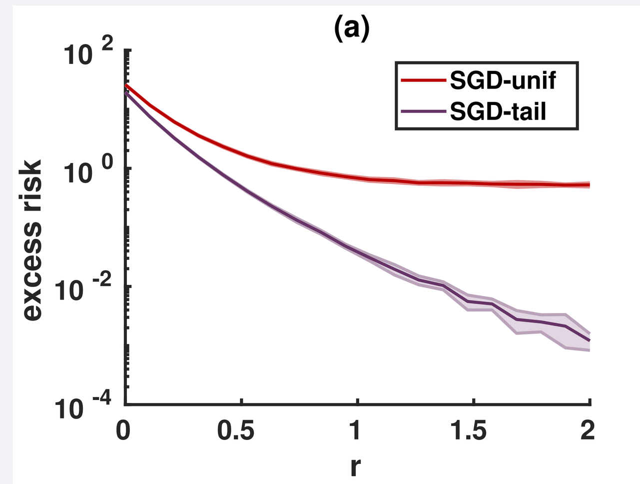
The excess risk of the (tail)-averaged SGD iterate satisfies

$$\mathbb{E}[\mathcal{L}(\bar{w}_L) - \mathcal{L}(w_*)] \lesssim n^{-\frac{2r+1}{2r+1+\nu}}$$

for each of the following choices:

- one pass: $b_n \simeq 1$, $L_n \simeq n$, $\gamma_n \simeq n^{-\frac{2r+\nu}{2r+1+\nu}}$
- one pass: $b_n \simeq n^{\frac{2r+\nu}{2r+1+\nu}}$, $L_n \simeq n^{\frac{1}{2r+1+\nu}}$, $\gamma_n \simeq 1$
- $\mathcal{O}\left(n^{\frac{1}{2r+1+\nu}}\right)$ passes: $b_n \simeq n$, $L_n \simeq n^{\frac{1}{2r+1+\nu}}$, $\gamma_n \simeq 1$.

Experiment: Saturation

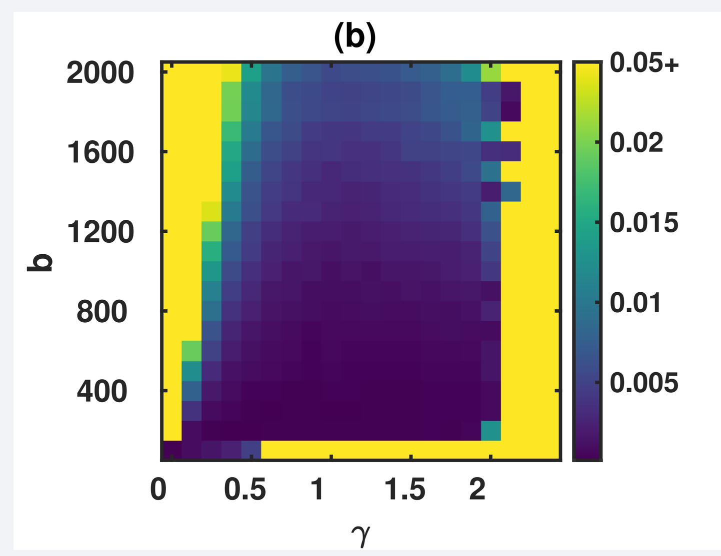


Excess risk as a function of regularity r with uniform and tail averaging.

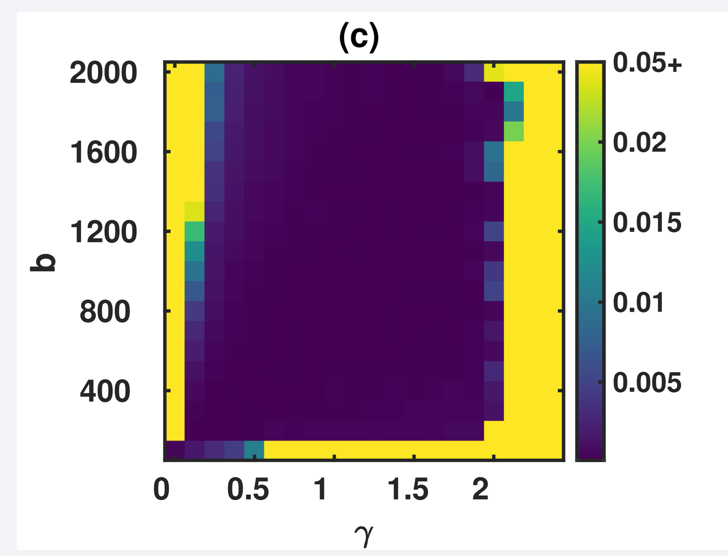
Unif AveSGD: starts to lag behind its tail-averaged counterpart for larger values of r exceeding $1/2$, flattening out.

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



Unif AveSGD



Tail AveSGD

Single pass performance as a function of the stepsize γ 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 STE 1074/4-1. L.R. acknowledges the financial support of the AFOSR projects FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), and the EU H2020-MSCA-RISE project NoMADS - DLV-777826.