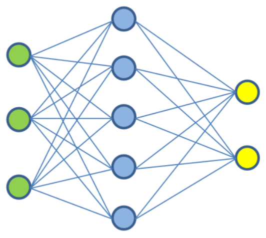
Last Iterate Risk Bounds of SGD with Decaying Stepsize for Overparameterized Linear Regression

Jingfeng Wu*[1], Difan Zou*[2], Vladimir Braverman[1], Quanquan Gu[2], Sham M. Kakade[3]

[1]Johns Hopkins University, [2]UCLA, [3]Harvard University

Population Risk

$\mathscr{L}(\mathbf{w}) = \mathbb{E}\ell(\mathbf{x}, y; \mathbf{w})$



***n* training samples**

$(\mathbf{x}_1, y_1)\cdots, (\mathbf{x}_n, y_n) \in \mathbb{R}^{d\times 1}$

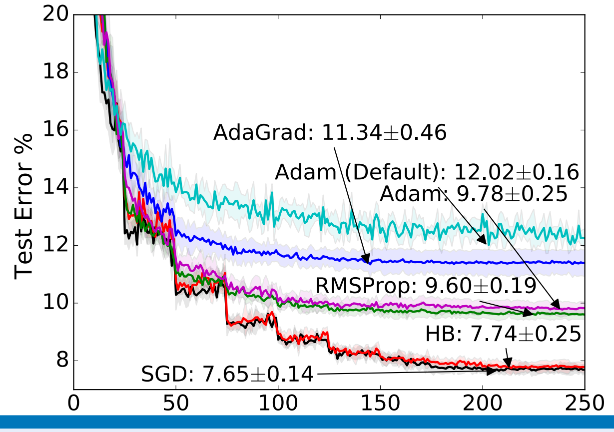
Large model

$\mathbf{w} \in \mathbb{R}^d$ for large d

SGD

$\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \nabla \ell(\mathbf{x}_i, y_i; \mathbf{w})$

SGD generalizes well (WRSSR 2017)



Least Square

True Model $\quad y = \mathbf{x}^\top \mathbf{w}^* + \mathcal{N}(0, \sigma^2)$

Data Covariance $\quad \mathbf{H} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top] =: \mathtt{diag}(\lambda_1, \lambda_2, \ldots)$, WOLG

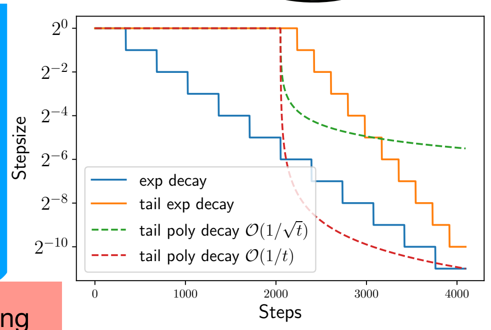
Population Risk $\quad \mathscr{L}(\mathbf{w}) := \mathbb{E}(y - \mathbf{x}^\top \mathbf{w})^2$

Excess Risk $\quad \Delta(\mathbf{w}) := \mathscr{L}(\mathbf{w}) - \mathscr{L}(\mathbf{w}^*) = (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^*)$

SGD

n samples $(\mathbf{x}_t, y_t)_{t=1}^n$

$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t \cdot (y_t - \mathbf{x}_t^\top \mathbf{w}_{t-1}) \cdot \mathbf{x}_t$ output $:= \mathbf{w}_n$



Stepsize Schedulers

[Strongly Contractive Fourth Moment Condition]

Recall that $\mathbf{H} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$. Assume that for every PSD matrix $\mathbf{A}$,

- $\mathbb{E}[\mathbf{x}^\top \mathbf{A} \mathbf{x} \cdot \mathbf{x}\mathbf{x}^\top] \preceq \alpha \cdot \mathrm{tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H}$ for some constant $\alpha \ge 1$;
- $\mathbb{E}[\mathbf{x}^\top \mathbf{A} \mathbf{x} \cdot \mathbf{x}\mathbf{x}^\top] \succeq \beta \cdot \mathrm{tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H} + \mathbf{H}\mathbf{A}\mathbf{H}$ for some constant $\beta > 0$.

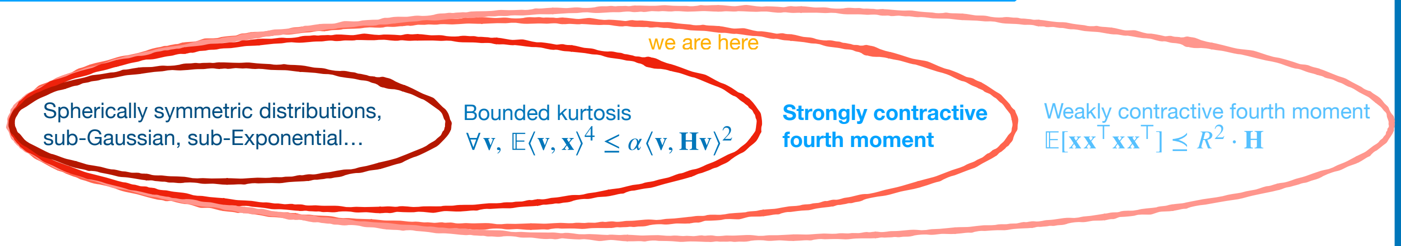
Geometrically Decaying

$$\eta_t = \begin{cases} \eta_0, & t \le s \\ 0.5\eta_{t-1}, & t > s, t \,\%\, K = 0 \\ \eta_{t-1}, & \text{otherwise} \end{cases}$$

Polynomially Decaying

$$\eta_t = \begin{cases} \eta_0, & t \le s \\ \dfrac{\eta_0}{(t-s)^a}, & t > s \end{cases}$$

for $0 \le a \le 1$



[WZBGK 2022]

Let $\mathbf{w}_n^{\mathtt{exp}}$ and $\mathbf{w}_n^{\mathtt{poly}}$ be the SGD outputs with geometrically and polynomially decaying stepsizes, respectively. Fix same $s = n/2$, same $\mathbf{w}_0$, same η_0. Then we have

$$\mathbb{E}\Delta(\mathbf{w}_n^{\mathtt{exp}}) \lesssim (1 + \mathtt{SNR} \cdot \log n) \cdot \mathbb{E}\Delta(\mathbf{w}_n^{\mathtt{poly}})$$

where $\mathtt{SNR} := \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 / \sigma^2$.

[WZBGK 2022]

Consider SGD with geometrically decaying stepsizes. Let the stepsize decaying interval be $K := (n-s)/\log(n-s)$. For every $s > 0$, $K > 2$ and every $\eta_0 < 1/(4\alpha\,\mathrm{tr}(\mathbf{H})\log(n))$, we have

$$\mathbb{E}\Delta(\mathbf{w}_n) \lesssim \frac{\|(\mathbf{I} - \eta_0\mathbf{H})^{s+K}(\mathbf{w}_0 - \mathbf{w}^*)\|_{\mathbf{I}_{0:k^*}}^2}{\eta_0 K} + \|(\mathbf{I} - \eta_0\mathbf{H})^{s+K}(\mathbf{w}_0 - \mathbf{w}^*)\|_{\mathbf{H}_{k^*:\infty}}^2$$
$$+ \frac{k^* + \eta_0 K \sum_{k^* < i \le k^\dagger} \lambda_i + \eta_0^2 K^2 \sum_{i > k^\dagger} \lambda_i^2}{K} \cdot \left(\sigma^2 + \alpha \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 \cdot \log(n)\right)$$

Here $k^*, k^\dagger$ are such that $\lambda_1 \ge \ldots \ge \lambda_{k^*} \ge \dfrac{1}{\eta_0 K} \ge \lambda_{k^*+1} \ge \ldots \ge \lambda_{k^\dagger} \ge \dfrac{1}{\eta_0(s+K)} \ge \lambda_{k^\dagger+1} \ge \ldots$

See the paper for a nearly matching lower bound.

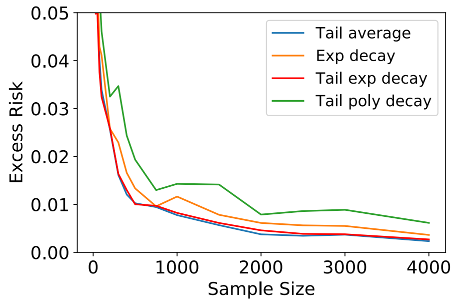
[GKKN 2019]

$$\mathbb{E}\Delta(\mathbf{w}_n) \lesssim \left(\frac{d\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\eta_0 n} + \frac{d}{n} \cdot \sigma^2 \right) \cdot \log n$$

Remarks

1. Weakly contractive fourth moment condition
2. Variance bound scales with d
3. ℓ_2-norm implicitly depends on d

Get the Paper!



Take Home

- Risk of SGD in high-dim $\approx d_{\mathtt{eff}} / n$
- $d_{\mathtt{eff}}$ determined by $(\lambda_i)_{i\ge 1}, \eta_0, n_{\mathtt{eff}}$; and $\ll d$ when $(\lambda_i)_{i\ge1}$ decay fast
- Geometrical stepsize > polynomially stepsize

JW is looking for Post Doc position!