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

Jingfeng Wu
with Difan Zou, Vladimir Braverman, Quanquan Gu, Sham M. Kakade

The Implicit Regularization Effect of SGD

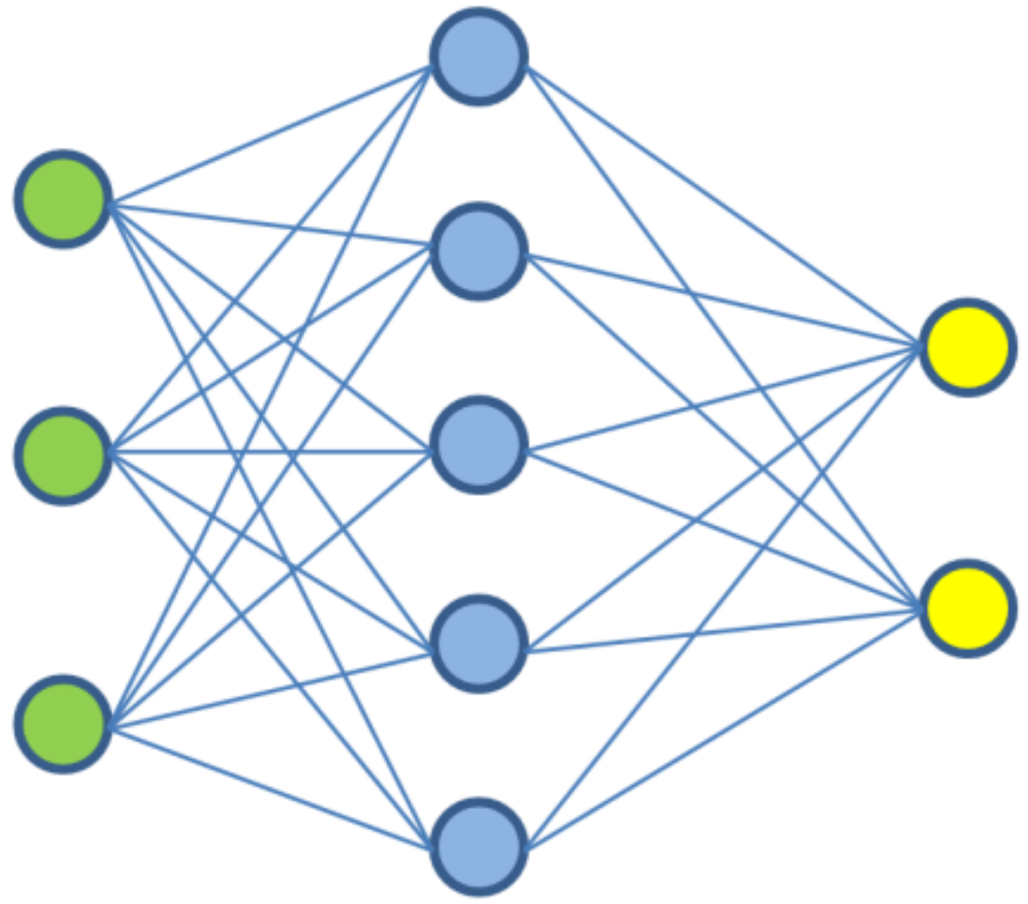
n **training samples**

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

Population Risk

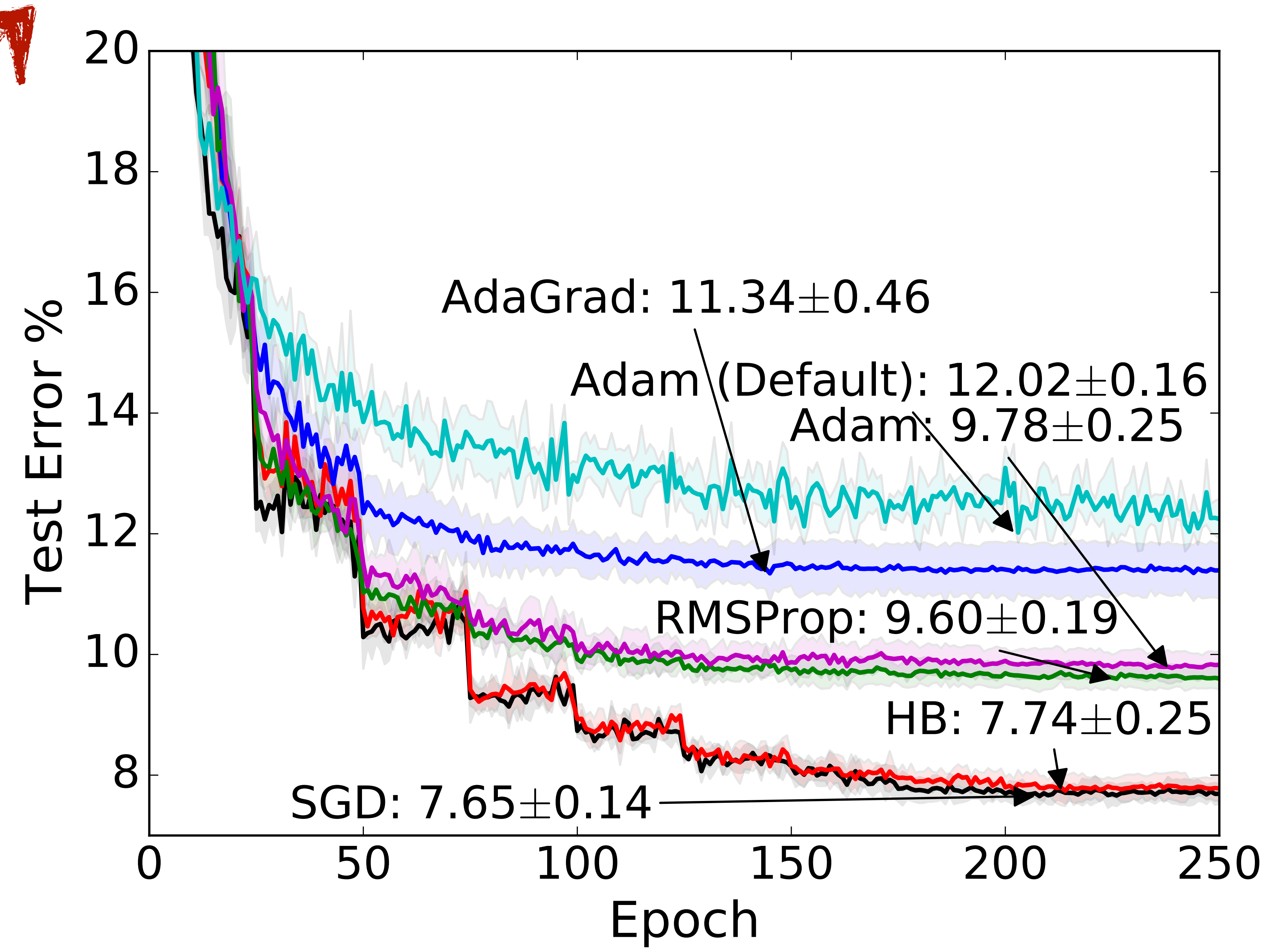
$\mathcal{L}(\mathbf{w}) = \mathbb{E}\ell(\mathbf{x}, y; \mathbf{w})$

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



Large Model

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



SGD generalizes well

SGD generalizes well for learning high-dim model

Wilson, Ashia C., Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. "The marginal value of adaptive gradient methods in machine learning." Advances in neural information processing systems 30 (2017).

High Dimensional Linear Regression

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

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

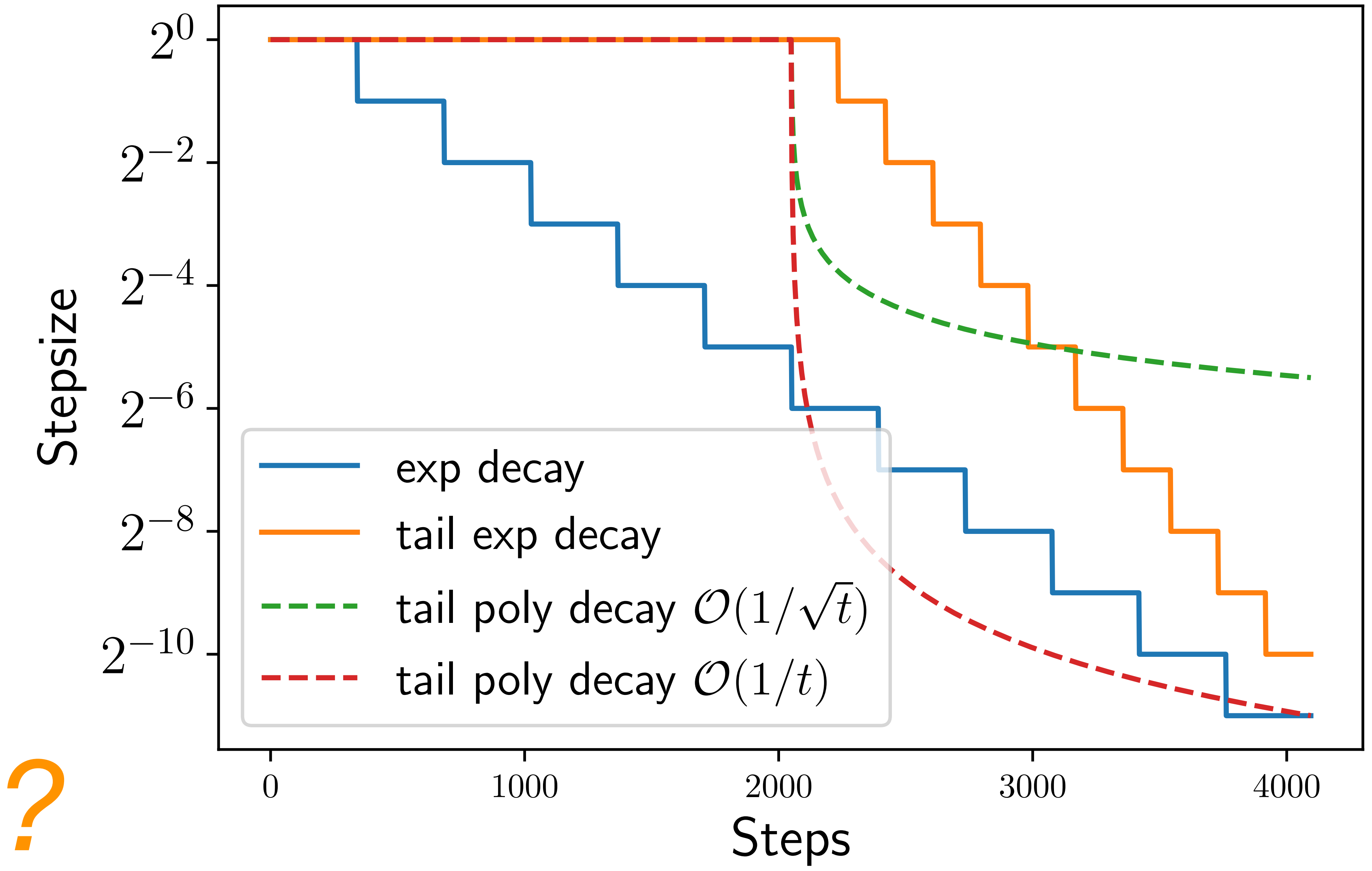
Population Risk $\mathcal{L}(\mathbf{w}) := \mathbb{E}(y - \mathbf{x}^\top \mathbf{w})^2$

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

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

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

output $:= \mathbf{w}_n$



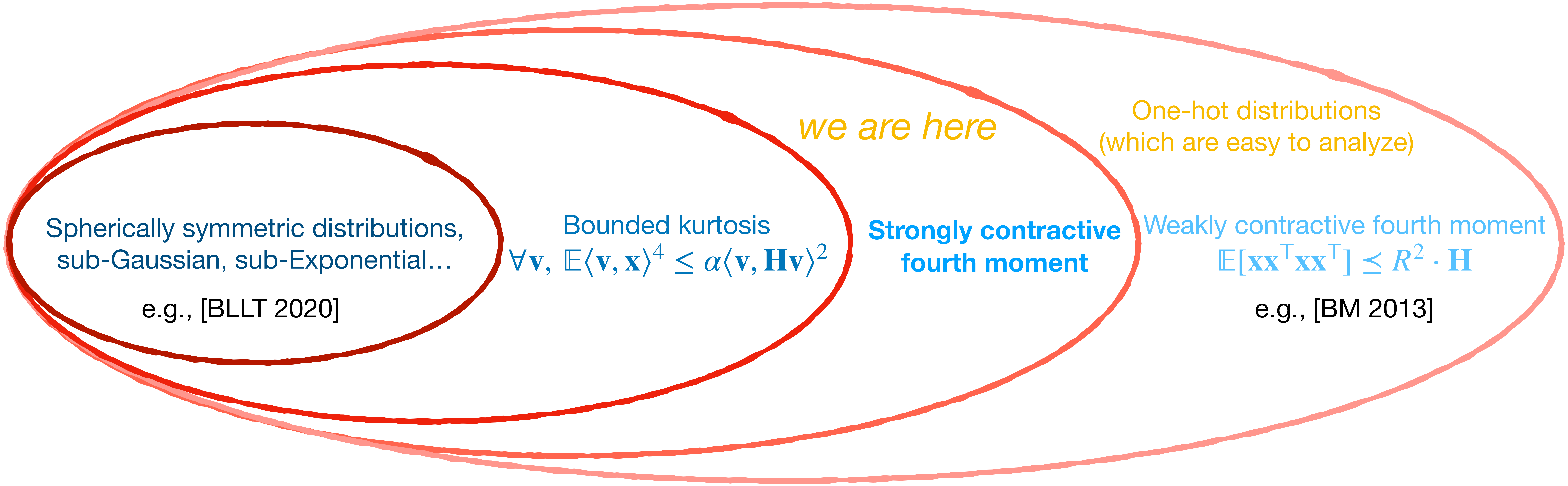
Caveat: One-Pass SGD

Two regimes: $d \lessgtr n$?

Key Assumption: Strongly Contractive Fourth Moment

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



- *Bach, Francis, and Eric Moulines. "Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n)." Advances in neural information processing systems 26 (2013).*
- *Bartlett, Peter L., Philip M. Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression." Proceedings of the National Academy of Sciences 117, no. 48 (2020): 30063-30070.*

Tail Geometrically Decaying Stepsizes

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

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

[GKKN 2019]

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

Useful in practice!

what if $d > n$ *?*

Remarks

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

Ge, Rong, Sham M. Kakade, Rahul Kidambi, and Praneeth Netrapalli. "The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares." Advances in Neural Information Processing Systems 32 (2019).

A Fine-Grained Upper Bound

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

exponentially decaying

$$\mathbb{E}\Delta(\mathbf{w}_n) \lesssim \frac{\|(\mathbf{I}-\eta_0\mathbf{H})^{s+K}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{I}_{0:k^*}}}{\eta_0 K} + \|(\mathbf{I}-\eta_0\mathbf{H})^{s+K}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{k^*:\infty}}$$

$$+ \frac{k^* + \eta_0 K \sum_{k^*<i\le k^\dagger}\lambda_i + \eta_0^2K^2\sum_{i>k^\dagger}\lambda_i^2}{K}\cdot\left(\sigma^2 + \alpha\cdot\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}}\cdot\log(n)\right)$$

effective dimension

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

Ambient Dimension d *vs.*

$\mathbf{I}_{0:k^*} :=$ diag$(1,\ldots,1,0,0,\ldots)$ $\quad \mathbf{H}_{k^*:\infty} :=$ diag$(0,\ldots,0,\lambda_{k^*+1},\lambda_{k^*+2},\ldots)$

Effective Dimension $k^* + \eta_0 K \sum_{k^*<i\le k^\dagger}\lambda_i + \eta_0^2K^2\sum_{i>k^\dagger}\lambda_i^2$, *small when* $(\lambda_i)_{i\ge1}$ *decays fast*

A Nearly Matching Lower Bound

Let the stepsize decaying interval be $K := (n-s)/\log(n-s)$. For every $s \geq 0$, $K > 10$ and every $\eta_0 < 1/\lambda_1$, we have

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

effective dimension

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

Lower bound nearly matches upper bound
if SNR is bounded, $\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}}^2 \lesssim \sigma^2$

$\mathbf{I}_{0:k^*} :=$ diag$(1,\ldots,1,0,0,\ldots)$
$\mathbf{H}_{k^*:\infty} :=$ diag$(0,\ldots,0,\lambda_{k^*+1},\lambda_{k^*+2},\ldots)$

Geometrically vs. Polynomially Decaying Stepsize

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

$$\eta_t = \begin{cases} \eta_0, & t \le s \\ \frac{\eta_0}{(t-s)^a}, & t > s \end{cases} \quad \text{for } 0 \le a \le 1$$

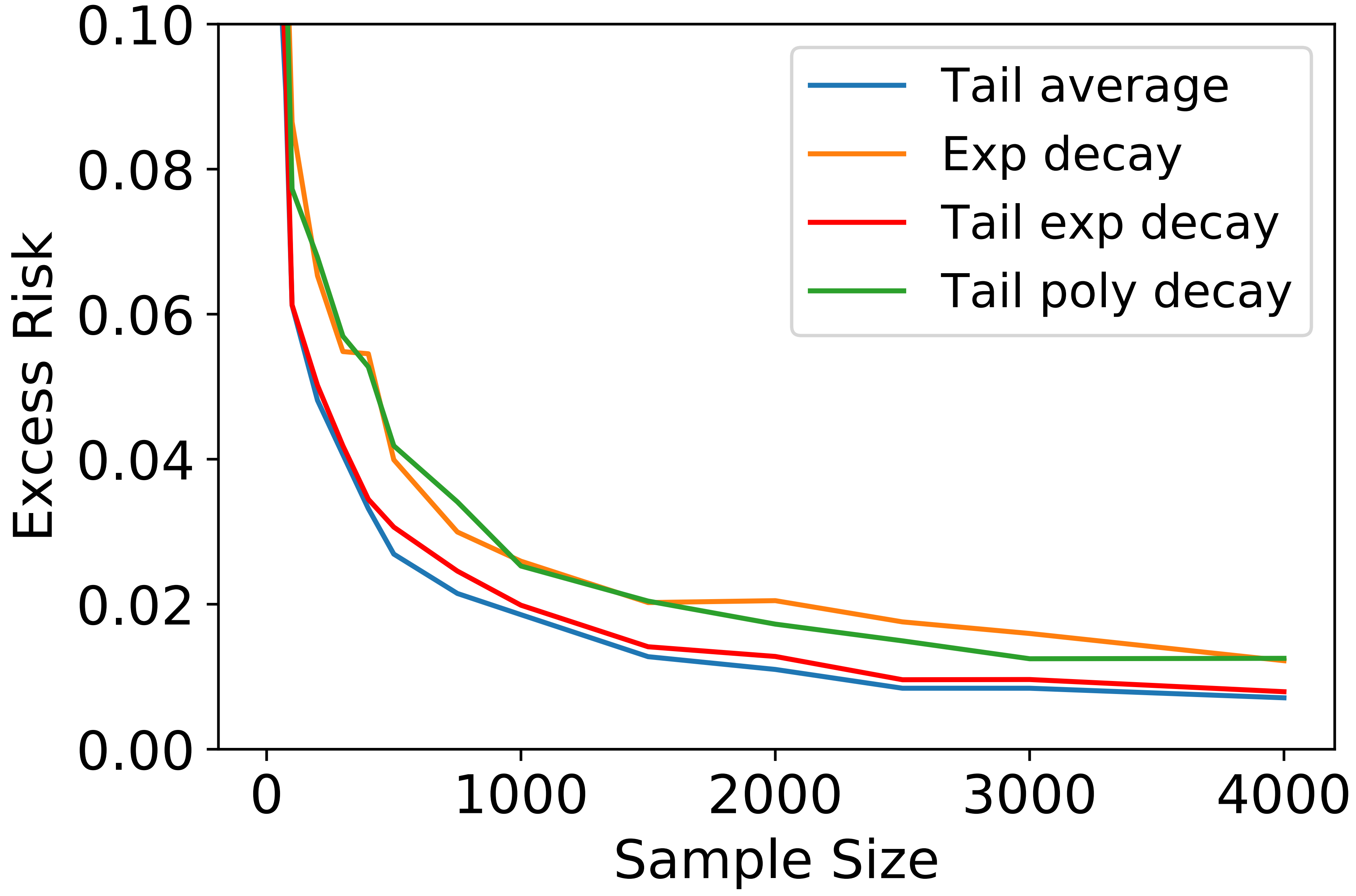
Let $\mathbf{w}_n^{\mathtt{exp}}$ and $\mathbf{w}_n^{\mathtt{poly}}$ be the SGD outputs with geometrically and polynomially decaying stepsizes, respectively. Fix same $s = n/2$, same $\mathbf{w}_0$, same η_0. Then we have

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

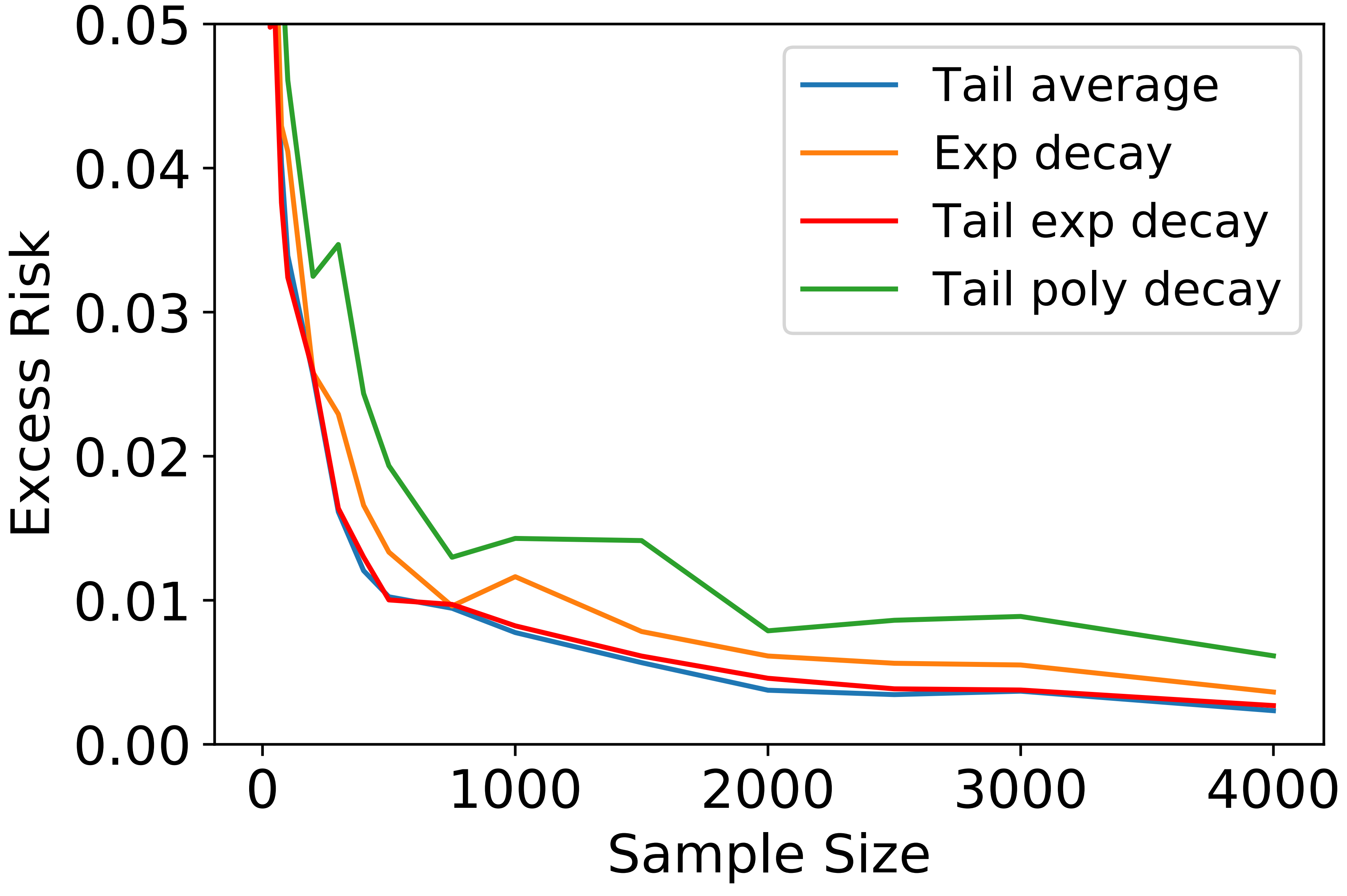
where $\mathrm{SNR} := \|\mathbf{w}_0 - \mathbf{w}_n\|_{\mathbf{H}}^2 \,/\, \sigma^2$.

*For **every** least square problem with bounded SNR,*
$\mathbf{w}_n^{\mathtt{exp}}$ is always nearly no worse than $\mathbf{w}_n^{\mathtt{poly}}$

Numerical Simulation



$\lambda_i = i^{-1}$, $\mathbf{w}^*[i] = i^{-1}$



$\lambda_i = i^{-2}$, $\mathbf{w}^*[i] = i^{-1}$

Experimental Setting: $\sigma^2 = 1$, $d = 256$, $\mathbf{w}_0 = 0$, $s = n/2$, $a = 1$

Under each sample size, the initial stepsize is fine-tuned for each algorithm

- SGD can generalize in high-dim least squares
- Geometrically decaying stepsizes > polynomially decaying stepsizes

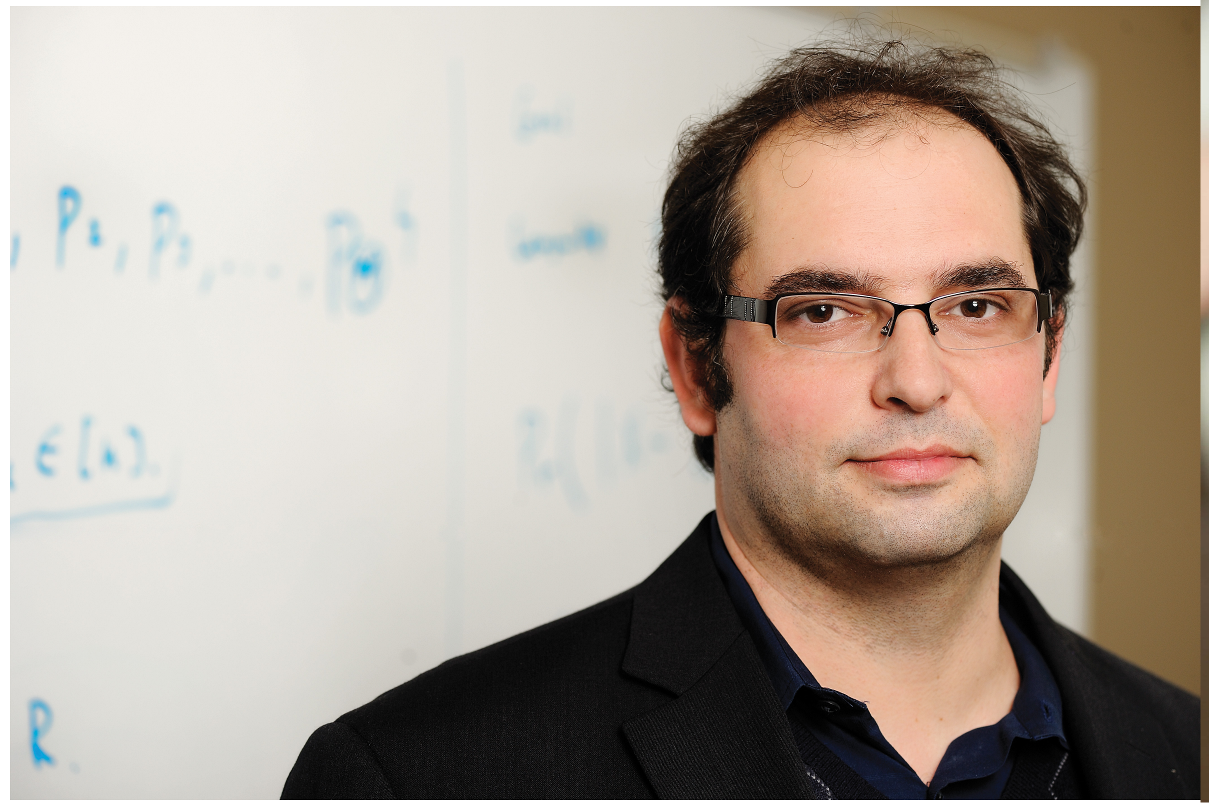
Conclusion

Take Home

- Risk of SGD in high-dim $\approx d_{\texttt{eff}} / n$
- $d_{\texttt{eff}}$ determined by $(\lambda_i)_{i\geq 1}$, η_0, $n_{\texttt{eff}}$; and $\ll d$ when $(\lambda_i)_{i\geq 1}$ decay fast
- Geometrical stepsize > polynomially stepsize

Limitations

- One-pass SGD
- Linear model
- Strongly contractive fourth moment



Vladimir Bravermen @ JHU

Quanquan Gu @ UCLA

Sham M. Kakade @ Harvard

Difan Zou @ UCLA

Get the Paper!