Benign Overfitting & Implicit Regularization of Constant-Stepsize SGD for Linear Regression

Jingfeng Wu

with Difan Zou, Vladimir Braverman, Quanquan Gu, Dean P. Foster, Sham M. Kakade

Outlines

1. Backgrounds
2. A Tight and Dimension-Free Bound for SGD
3. The Benign-Overfitting Phenomenon
4. SGD vs. Ridge Regression: Implicit vs. Explicit Regularization
5. Proof Sketches

Linear Regression Problems

A Problem Instance

- Label: $y = \langle \mathbf{w}^*, \mathbf{x} \rangle + \xi$
- Noise: $\mathbb{E}[\xi] = 0$, $\mathbb{E}[\xi^2] = \sigma^2$, $\xi \perp \mathbf{x}$ can be relaxed
- Data: $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{H} = \text{diag}(\lambda_1, \lambda_2, \ldots)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$
- Risk: $L_D(\mathbf{w}) := \mathbb{E}(y - \langle \mathbf{w}, \mathbf{x} \rangle)^2$
- Goal: Minimizing ExcessRisk$(\mathbf{w}) := L_D(\mathbf{w}) - L_D(\mathbf{w}^*) = \|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{H}}^2$
- Information: $(\mathbf{x}_1, y_1) \cdots, (\mathbf{x}_N, y_N) \in \mathbb{R}^d$, i.i.d.

Notations

- $\|\mathbf{v}\|_{\mathbf{H}}^2 := \mathbf{v}^\top \mathbf{H} \mathbf{v}$,
- $\mathbf{H}_{0:k} := \text{diag}(\lambda_1, \ldots, \lambda_k, 0, \ldots)$
- $\mathbf{H}_{k:\infty} := \text{diag}(0, \ldots, 0, \lambda_{k+1}, \lambda_{k+2}, \ldots)$

Two regimes: $d \lessgtr N$

Data Distributions

Assumption [The fourth-order moment tensor]

- For every PSD matrix $\mathbf{A}$, $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A} \mathbf{x}\mathbf{x}^\top] \preceq \alpha \cdot \mathrm{Tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H}$ for some constant α.
- For every PSD matrix $\mathbf{A}$, $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A} \mathbf{x}\mathbf{x}^\top] \succeq \beta \cdot \mathrm{Tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H}$ for some constant β.

Remark

1. Assumption holds for every *sub-Gaussian* distributions, e.g., for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$, we can set $\alpha = 3$ and $\beta = 2$.
2. Assumption only needs to hold for $\mathbf{A}$ that commutes with $\mathbf{H}$.

Constant-Stepsize SGD with Iterate-Averaging

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma \cdot (y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t \rangle)\mathbf{x}_t \qquad \text{output} := \frac{1}{N}\sum_{t=0}^{N-1} \mathbf{w}_t$$

[BM 2013]

$$\mathbb{E}[\text{ExcessRisk}] \lesssim \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N} + \frac{d}{N} \cdot \sigma^2$$

[JNKKS 2018]

$$\mathbb{E}[\text{ExcessRisk}] \lesssim \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2}{\lambda_{\min}^2 \gamma^2 N^2} + \frac{d}{N} \cdot \sigma^2$$

Remarks

1. Minimax optimal when $d < N$
2. Variance bound scales with d
3. ℓ_2-norm or condition number implicitly depends on d

What if $d > N$??

Outlines

1. Backgrounds
2. A Tight and Dimension-Free Bound for SGD
3. The Benign-Overfitting Phenomenon
4. SGD vs. Ridge Regression: Implicit vs. Explicit Regularization
5. Proof Sketches

An Upper Bound: Dimension-Free

For every N, every $\gamma < 1/(2\alpha \mathrm{Tr}(\mathbf{H}))$ and every k^*, we have

$$\mathbb{E}[\text{ExcessRisk}] \lesssim \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}}}{\gamma^2 N^2} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} +$$

$$\frac{k^* + \gamma^2 N^2 \sum_{i>k^*} \lambda_i^2}{N} \cdot \left(\sigma^2 + \alpha \left(\frac{\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}}}{\gamma N} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{0:k^*}} \right) \right)$$

In particular, k^* could be such that $\lambda_1 \geq \ldots \geq \lambda_{k^*} \geq \frac{1}{\gamma N} \geq \lambda_{k^*+1} \geq \ldots$

Algorithm

learnable dimensions $\lambda_1 \geq \ldots \geq \lambda_{k^*} \geq \frac{1}{\gamma N} \geq \lambda_{k^*+1} \geq \ldots$ unlearnable dimensions

Data

A Lower Bound: Instance-Wisely Tight

For every sufficiently large N and every $\gamma < 1/\lambda_1$, we have

$$\mathbb{E}[\text{ExcessRisk}] \gtrsim \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}}}{\gamma^2 N^2} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} +$$

$$\frac{k^* + \gamma^2 N^2 \sum_{i>k^*} \lambda_i^2}{N} \cdot \left(\sigma^2 + \beta \left(\frac{\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}}}{\gamma N} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{0:k^*}} \right) \right)$$

Here k^* is such that $\lambda_1 \geq \ldots \geq \lambda_{k^*} \geq \frac{1}{\gamma N} \geq \lambda_{k^*+1} \geq \ldots$

Algorithm

learnable dimensions $\lambda_1 \geq \ldots \geq \lambda_{k^*} \geq \frac{1}{\gamma N} \geq \lambda_{k^*+1} \geq \ldots$ unlearnable dimensions

Data

A Dimension-Free Bound

	Ours	[BM 2013]
Bias	$\frac{\lVert \mathbf{w}_0 - \mathbf{w}^* \rVert^2_{\mathbf{H}^{-1}_{0:k^*}}}{\gamma^2 N^2} + \lVert \mathbf{w}_0 - \mathbf{w}^* \rVert^2_{\mathbf{H}_{k^*:\infty}}$	$\frac{\lVert \mathbf{w}_0 - \mathbf{w}^* \rVert_2^2}{\gamma N}$
"Dimension"	$k^* + \gamma^2 N^2 \sum_{i>k^*} \lambda_i^2$	d
Noise	$\sigma^2 + \alpha \left(\frac{\lVert \mathbf{w}_0 - \mathbf{w}^* \rVert^2_{\mathbf{I}_{0:k^*}}}{\gamma N} + \lVert \mathbf{w}_0 - \mathbf{w}^* \rVert^2_{\mathbf{H}_{0:k^*}} \right)$	σ^2

Even when $d > N$, SGD can generalize if the spectrum decays fast

Examples

Suppose that $\|\mathbf{w}_0 - \mathbf{w}^*\|_2 \leq \mathcal{O}(1)$, $\sigma^2 \leq \mathcal{O}(1)$, and $\gamma = 1/(2\alpha \mathrm{Tr}(\mathbf{H}))$:

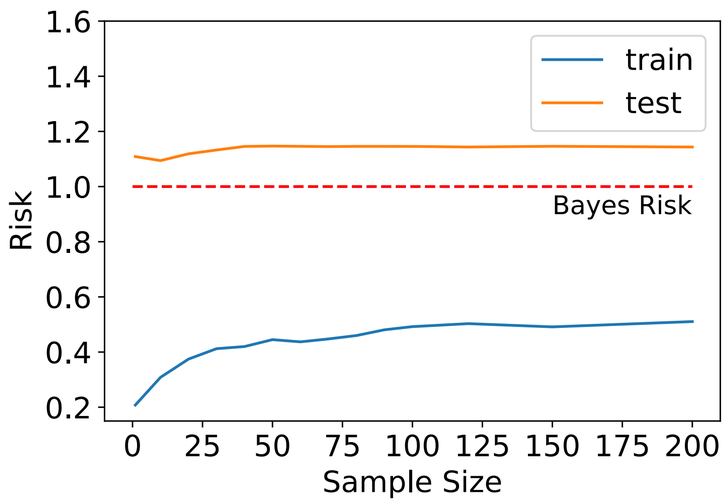
- for $\lambda_i = i^{-(1+r)}$, $r > 0$, $\mathbb{E}[\text{ExcessRisk}] \leq \mathcal{O}\left(N^{-r/(1+r)}\right)$;
- for $\lambda_i = i^{-1}\log^{-s}(i+1)$, $s > 1$, $\mathbb{E}[\text{ExcessRisk}] \leq \mathcal{O}\left(\log^{-s}(N)\right)$;
- For $\lambda_i = e^{-i}$, $\mathbb{E}[\text{ExcessRisk}] \leq \mathcal{O}\left(\log(N)/N\right)$.

Even when $d > N$, SGD can generalize if the spectrum decays fast

Outlines

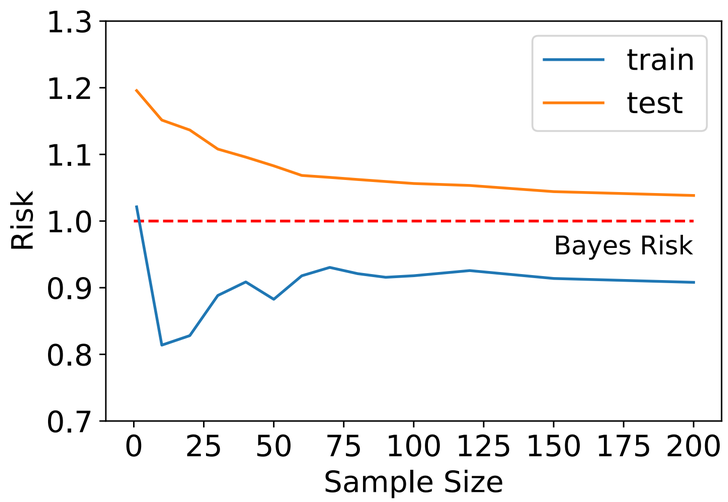
1. Backgrounds
2. A Tight and Dimension-Free Bound for SGD
3. The Benign-Overfitting Phenomenon
4. SGD vs. Ridge Regression: Implicit vs. Explicit Regularization
5. Proof Sketches

The Benign-Overfitting Phenomenon



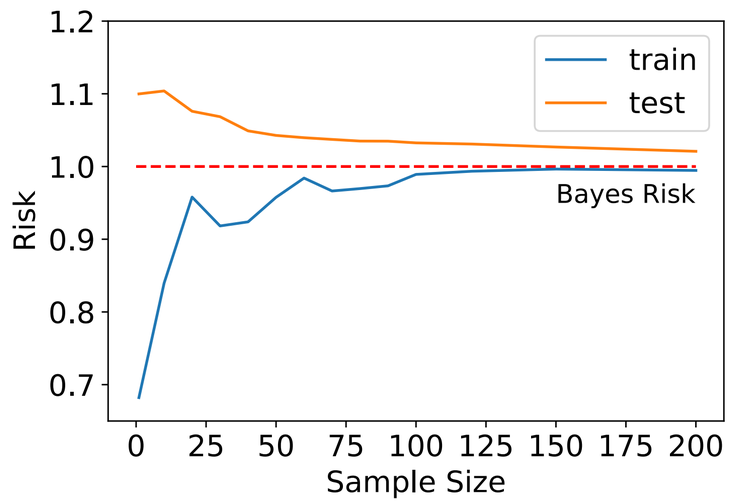
(a) $\lambda_i = i^{-1}$**, overfitting**

Overfits training data ✓
Generalizes on test data ✗



(b) $\lambda_i = i^{-1}\log^{-2}(i)$**, benign overfitting**

Overfits training data ✓
Generalizes on test data ✓



(c) $\lambda_i = i^{-2}$**, regularization**

Overfits training data ✗
Generalizes on test data ✓

Settings:

$\sigma = 1$

$d = 2{,}000$

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

Conjecture: benign overfitting for SGD happens when

$\lambda_i = i^{-r}\log^{-s}(i)$ for $r = 1$ and $s > 1$

Evidence: This is true for Ordinary Least Square [BLLT 2020]

Outlines

1. Backgrounds
2. A Tight and Dimension-Free Bound for SGD
3. The Benign-Overfitting Phenomenon
4. SGD vs. Ridge Regression: Implicit vs. Explicit Regularization
5. Proof Sketches

Constant-Stepsize SGD with Tail-Averaging

$$\text{output} := \frac{2}{N}\sum_{t=N/2}^{N-1} \mathbf{w}_t$$

Geometric improvement for the bias error in the head space

For every N, every $\gamma < 1/(2\alpha \mathrm{Tr}(\mathbf{H}))$ and every k^*, we have

$$\mathbb{E}[\text{ExcessRisk}] \lesssim \frac{\|(\mathbf{I}-\gamma\mathbf{H})^{N/2}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{0:k^*}}}{\gamma^2 N^2} + \|(\mathbf{I}-\gamma\mathbf{H})^{N/2}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{k^*:\infty}} +$$

$$\frac{k^* + \gamma^2 N^2 \sum_{i>k^*}\lambda_i^2}{N} \cdot \left(\sigma^2 + \alpha\left(\boxed{\frac{\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}}}{\gamma N}} + \|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{0:k^*}}\right)\right)$$

Open Problem: tight or not?

In particular, k^* could be such that $\lambda_1 \geq \ldots \geq \lambda_{k^*} \geq \frac{1}{\gamma N} \geq \lambda_{k^*+1} \geq \ldots$

Ridge Regression vs. SGD

- Ridge

$$\text{output} := \arg\min_{\mathbf{w}} \sum_{t=0}^{N-1} \|\mathbf{x}_t^\top \mathbf{w} - y_t\|_2^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

$$= \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Algorithmic, tunable hyperparameters

- SGD (tail-averaging + zero initialization)

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma \cdot (y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t \rangle)\mathbf{x}_t, \text{ output } := \frac{2}{N} \sum_{t=N/2}^{N-1} \mathbf{w}_t,$$

$$\gamma < 1/(2\alpha \mathrm{Tr}(\mathbf{H}))$$

Ridge Regression

[TB 2020, ZWBGFK 2021] Suppose the data distribution is symmetric. For every sufficiently large N and every λ, we have

$$\mathbb{E}[\text{ExcessRisk}] \gtrsim \frac{(\lambda + \sum_{i>k^*} \lambda_i)^2}{N^2} \cdot \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} + \left(k^* + \frac{N^2}{(\lambda + \sum_{i>k^*} \lambda_i)^2} \sum_{i>k^*} \lambda_i^2 \right) \cdot \frac{\sigma^2}{N}$$

Here k^* is such that $\lambda_1 \geq \ldots \geq \lambda_{k^*} \geq \frac{\lambda + \sum_{i>k^*} \lambda_i}{N} \geq \lambda_{k^*+1} \geq \ldots$

Algorithm

learnable dimensions $\lambda_1 \geq \ldots \geq \lambda_{k^*} \geq \frac{\lambda + \sum_{i>k^*} \lambda_i}{N} \geq \lambda_{k^*+1} \geq \cdots$ unlearnable dimensions

Data

Algorithm-Dependent Effective Dimension

	SGD (tail-averaging, zero initialization)	Ridge Regularization [TB 2020]
Bias	$\frac{\|(\mathbf{I}-\gamma\mathbf{H})^{N/2}\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}}}{\gamma^2 N^2} + \|(\mathbf{I}-\gamma\mathbf{H})^{N/2}\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}$	$\frac{(\lambda+\sum_{i>k^*}\lambda_i)^2}{N^2}\cdot\|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}$
"Dimension"	$k^* + \gamma^2 N^2 \sum_{i>k^*}\lambda_i^2$	$k^* + \frac{N^2}{(\lambda+\sum_{i>k^*}\lambda_i)^2}\sum_{i>k^*}\lambda_i^2$
Noise	$\sigma^2 + \alpha\left(\frac{\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}}}{\gamma N} + \|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{0:k^*}}\right)$	σ^2
Optimal k^*	$\lambda_1 \geq \ldots \geq \lambda_{k^*} \approx \frac{1}{\gamma N} \geq \ldots$	$\lambda_1 \geq \ldots \geq \lambda_{k^*} \approx \frac{\lambda+\sum_{i>k^*}\lambda_i}{N} \geq \ldots$

Ridge with large λ corresponds to SGD with small γ

Implicit vs. Explicit Regularization

Consider "well-tuned" SGD (tail-averaging) and Ridge, and *a class of Gaussian least square problems*

$$\mathcal{G} := \left\{\mathbf{w}^*, \mathbf{H}, \sigma^2 : \mathbf{x} \sim \mathcal{N}(0, \mathbf{H}),\ y = \mathbf{x}^\top \mathbf{w}^* + \xi,\ \xi \sim \mathcal{N}(0, \sigma^2)\right\}$$

- For every problem in $\mathcal{G}$, we have $\mathbb{E}[\text{ExcessRisk}_{\text{sgd}}] \lesssim \mathbb{E}[\text{ExcessRisk}_{\text{ridge}}]$ if $N_{\text{sgd}} \geq (1 + R^2) \cdot \kappa(N_{\text{ridge}}) \cdot \log(a) \cdot N_{\text{ridge}}$
- There is a problem in $\mathcal{G}$ with $R^2 = 1$, $\kappa(N_{\text{sgd}}) \approx 1$, such that for $\mathbb{E}[\text{ExcessRisk}_{\text{ridge}}] \lesssim \mathbb{E}[\text{ExcessRisk}_{\text{sgd}}]$ we must have $N_{\text{ridge}} \geq N_{\text{sgd}}^2 / \log^2(N_{\text{sgd}})$

$$R^2 := \frac{\|\mathbf{w}\|_{\mathbf{H}}^2}{\sigma^2}, \quad \kappa(n) := \frac{\text{Tr}(\mathbf{H})}{n\lambda_{\min\{n,d\}}}, \quad a := \kappa(N_{\text{ridge}}) R \sqrt{N_{\text{ridge}}}$$

Ridge could be bad ☹️

SGD is nearly alway good 🙂

Outlines

1. Backgrounds
2. A Tight and Dimension-Free Bound for SGD
3. The Benign-Overfitting Phenomenon
4. SGD vs. Ridge Regression: Implicit vs. Explicit Regularization
5. Proof Sketches

Proof Sketch: Bias-Variance Decomposition

Bias-Variance Decomposition in the PSD Matrix Space

$$\mathbb{E}[\text{ExcessRisk}] \lesssim \frac{1}{\gamma N^2}\langle \mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^N, \sum_{t=0}^{N-1} \mathbf{B}_t \rangle + \frac{1}{\gamma N^2}\langle \mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^N, \sum_{t=0}^{N-1} \mathbf{C}_t \rangle$$

SGD Iterate = Bias Iterate + Variance Iterate, i.e., $\mathbb{E}[(\mathbf{w}_t - \mathbf{w}^*)(\mathbf{w}_t - \mathbf{w}^*)^\top] = \mathbf{B}_t + \mathbf{C}_t \in \mathbb{R}^{d \times d}$

Bias Iterate: set noise to zero

$$\begin{cases} \mathbf{B}_t = (\mathfrak{I} - \gamma \mathfrak{T}) \circ \mathbf{B}_t \\ \mathbf{B}_0 = (\mathbf{w}_0 - \mathbf{w}^*)(\mathbf{w}_0 - \mathbf{w}^*)^\top \end{cases}$$

Variance Iterate: set initialization to the optimal

$$\begin{cases} \mathbf{C}_t = (\mathfrak{I} - \gamma \mathfrak{T}) \circ \mathbf{C}_{t-1} + \gamma^2 \sigma^2 \cdot \mathbf{H} \\ \mathbf{C}_0 = 0 \end{cases}$$

Tensor Operators

$$\mathfrak{I} := \mathbf{I} \otimes \mathbf{I}$$

$$\mathfrak{M} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top \otimes \mathbf{x}\mathbf{x}^\top]$$

$$\mathfrak{T} := \mathbf{I} \otimes \mathbf{H} + \mathbf{H} \otimes \mathbf{I} - \gamma \mathfrak{M}$$

SGD update rule

$$(\mathfrak{I} - \gamma \mathfrak{T}) \circ A = \mathbb{E}[(\mathbf{I} - \gamma \mathbf{x}\mathbf{x}^\top) A (\mathbf{I} - \gamma \mathbf{x}\mathbf{x}^\top)]$$

Assumption: $\mathfrak{M} \circ \mathbf{A} \preceq \alpha \cdot \mathrm{Tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H}$

Proof Sketch: Control Variance

A Crude Upper Bound on $\mathbf{C}_t$

For every t, we have $\mathbf{C}_t \preceq \dfrac{\gamma^2\sigma^2}{1-\gamma\alpha\mathrm{Tr}(\mathbf{H})} \cdot \mathbf{I}\left(=\mathbf{C}_\infty\right)$.

easily verifiable by induction

A Refined Bound on $\mathbf{C}_t$

$$
\begin{aligned}
\mathbf{C}_t &= (\mathfrak{I} - \gamma\mathfrak{T}) \circ \mathbf{C}_{t-1} + \gamma^2\sigma^2 \cdot \mathbf{H} \\
&\preceq (\mathbf{I}-\gamma\mathbf{H})\mathbf{C}_{t-1}(\mathbf{I}-\gamma\mathbf{H}) + \gamma^2\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{C}_{t-1}\mathbf{x}\mathbf{x}^\top] + \gamma^2\sigma^2 \cdot \mathbf{H} \\
&\preceq (\mathbf{I}-\gamma\mathbf{H})\mathbf{C}_{t-1}(\mathbf{I}-\gamma\mathbf{H}) + \gamma^2\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{C}_{\infty}\mathbf{x}\mathbf{x}^\top] + \gamma^2\sigma^2 \cdot \mathbf{H} \\
&\preceq (\mathbf{I}-\gamma\mathbf{H})\mathbf{C}_{t-1}(\mathbf{I}-\gamma\mathbf{H}) + \frac{\gamma^2\sigma^2}{1-\gamma\alpha\mathrm{Tr}(\mathbf{H})} \cdot \mathbf{H}
\end{aligned}
$$

$$\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N \preceq \mathbf{I}_{0:k^*} + \gamma N \cdot \mathbf{H}_{k^*:\infty}$$

$$\mathbf{C}_t \preceq \frac{\gamma\sigma^2}{1-\gamma\alpha\mathrm{Tr}(\mathbf{H})} \cdot \left(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\right)$$

Proof Sketch: Control Bias Summations

A Crude Upper Bound on $\mathbf{S}_t := \sum_{n=0}^{t-1} \mathbf{B}_t$

For every t, we have $\mathbf{S}_t \preceq \frac{1}{\gamma} \cdot \mathfrak{T}^{-1} \circ \left((\mathbf{I} - \gamma \mathbf{H})^N \mathbf{B}_0 (\mathbf{I} - \gamma \mathbf{H})^N \right)$.

A Refined Bound on $\mathbf{S}_t$

$$
\begin{aligned}
\mathbf{S}_t &= (\mathfrak{I} - \gamma \mathfrak{T}) \circ \mathbf{S}_{t-1} + \mathbf{B}_0 \\
&\preceq (\mathbf{I} - \gamma \mathbf{H}) \mathbf{S}_{t-1} (\mathbf{I} - \gamma \mathbf{H}) + \gamma^2 \mathfrak{M} \circ \mathbf{S}_{t-1} + \mathbf{B}_0 \\
&\preceq (\mathbf{I} - \gamma \mathbf{H}) \mathbf{S}_{t-1} (\mathbf{I} - \gamma \mathbf{H}) + \gamma \mathfrak{M} \circ \mathfrak{T}^{-1} \circ \left((\mathbf{I} - \gamma \mathbf{H})^N \mathbf{B}_0 (\mathbf{I} - \gamma \mathbf{H})^N \right) + \mathbf{B}_0
\end{aligned}
$$

solve this recursion

An Important Lemma

$$\mathfrak{M} \circ \mathfrak{T}^{-1} \circ \mathbf{A} \preceq \frac{\alpha \mathrm{Tr}(\mathbf{A})}{1 - \alpha \gamma \mathrm{Tr} \mathbf{H}} \cdot \mathbf{H}$$

Open Problem: tight bound on $\mathbf{B}_t$?

Take Home

- Rethinking a problem “dimension”: algorithm matters.
- SGD can *benign overfit*.
- SGD has *implicit regularization* effect, that nearly *dominates* the explicit, ridge regularization effect.
- Tail-averaging improves the SGD bias in the head space.

References

1. Bach, Francis, and Eric Moulines. "Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n)." arXiv preprint arXiv:1306.2119 (2013).

2. 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.

3. Jain, Prateek, Sham Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. "Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification." Journal of Machine Learning Research 18 (2018).

4. Tsigler, Alexander, and Peter L. Bartlett. "Benign overfitting in ridge regression." arXiv preprint arXiv:2009.14286 (2020).

5. Zou, Difan*, Jingfeng Wu*, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. "Benign overfitting of constant-stepsize sgd for linear regression." COLT (2021).

6. Zou, Difan*, Jingfeng Wu*, Vladimir Braverman, Quanquan Gu, Dean P. Foster, and Sham M. Kakade. "The Benefits of Implicit Regularization from SGD in Least Squares Problems." arXiv preprint arXiv:2108.04552 (2021).