Nearly Optimal Bounds for Cyclic Forgetting

Halyun Jeong[1] Mark Kong[1] Deanna Needell[1]
William Swartworth[2] Rachel Ward[3]

[1]University of California, Los Angeles

[2]Carnegie Mellon University

[3]University of Texas at Austin

Overview: Our 2 main results

1. Numerical range bound: Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Let $\mathcal{P}_d = \{\text{orthogonal projections in } \mathbb{F}^d\}$ and let $\mathcal{P}_d^T = \prod_{t=1}^T \mathcal{P}_d$ be the Minkowski product. Then $\bigcup_{\substack{A \in \mathcal{P}_d^T \\ d \in \mathbb{Z}_{\geq 1}}} W(A)$ is a (closed, filled-in) sinusoidal spiral.
   - Corollary: Let $A \in \mathcal{P}_d^T$. Then $\|A^m(1-A)\| = O(\frac{T}{m})$.
2. Improved bound on forgetting in continual learning
   - T suitably normalized datasets, at least one of rank $r_{\max}$, in $\mathbb{R}^d$, cycled through m times
   - Trivial bound: 1
   - Known lower bound[1]: $\Omega(\frac{T^2}{mT}) = \Omega(\frac{T}{m})$
   - Old upper bound[2]: $\left\{ \frac{T^2}{\sqrt{mT}}, \frac{T^2(d - r_{\max})}{2mT} \right\}$
   - New upper bound: $O(\frac{T^2}{m})$ with reasonable constant[3]

---

[1] Itay Evron et al. "How catastrophic can catastrophic forgetting be in linear regression?" In: *Conference on Learning Theory*. PMLR. 2022, pp. 4028–4079.

[2] Ibid.

[3] Plus minor constant-factor optimizations not due to the numerical range bound

Review of Forgetting

- Continual learning: An ML algorithm (with parameters initialized at $\vec{w}_0$) is given a sequence S of tasks to learn over, with corresponding loss functions $\mathcal{L}_1, \mathcal{L}_2, \ldots,$ yielding parameter vectors $\vec{w}_1, \vec{w}_2, \ldots$ after each task
- Assuming[4] $\mathcal{L}_t(\vec{w}_t) = 0$ for all t, forgetting after nth update is

$$F_S(n) := \frac{1}{n} \sum_{t=1}^{n} \mathcal{L}_t(\vec{w}_n)$$

- In words: Average loss over all previously seen tasks, evaluated at nth learned parameter vector.
- Each task is weighted equally, but our results generalize to weighted forgetting with weights $W_1, W_2, \dots \in \mathbb{R}$:

$$\frac{1}{n} \sum_{t=1}^{n} W_t \mathcal{L}_t(\vec{w}_n)$$

[4] Relaxed slightly in paper

Our Setting

- Tasks are linear regression over datasets $(X_t, \vec{y}_t)$
- Loss is sum of squares error
- Datasets visited cyclically, in cycles of length T: $(X_1, \vec{y}_1), (X_2, \vec{y}_2), \ldots, (X_T, \vec{y}_T), (X_1, \vec{y}_1), \ldots$
- Datasets jointly realizable[5]
- Learning algorithm orthogonally projects onto solution space at each step
- Only consider forgetting after a whole number of cycles[6], so forgetting becomes average loss over all datasets

[5] Can be relaxed slightly
[6] Can be relaxed via weighted forgetting

Our Approach

Recall $\mathcal{P}_d = \{\text{orthogonal projections in } \mathbb{F}^d\}$. Take $\mathbb{F} = \mathbb{R}$.
Previously known bounds:

- If the datasets $X_1, \ldots, X_T \subset \mathbb{R}^d$ are normalized[7] so $\max_t \|X_t\| \leq 1$ and other suitable normalizations hold, then[8]

$$F_S(mT) \leq \frac{T-1}{2} \max_{A \in \mathcal{P}_d^T} \|A^m(1-A)\|$$

- There exists $Q \in \mathbb{R}$ such that, for any complex Hilbert space H, any linear map $\varphi : H \to H$, and any polynomial $f \in \mathbb{C}[z]$,

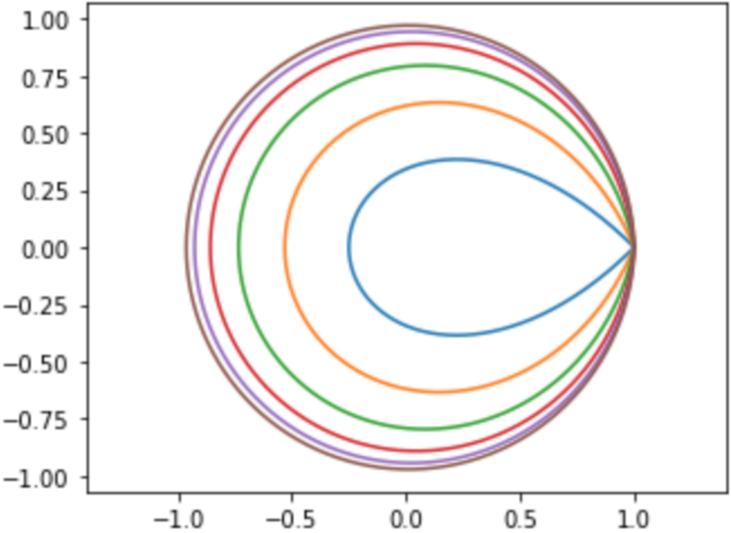
$$\|f(\varphi)\| \leq Q \sup_{z \in W(\varphi)} |f(z)|.$$

  Best known[9] value of Q is $1 + \sqrt{2}$.

[7] Alternative normalizations are possible

[8] Evron et al., "How catastrophic can catastrophic forgetting be in linear regression?"

[9] Michel Crouzeix and César Palencia. "The numerical range is a (1+2)-spectral set". In: *SIAM Journal on Matrix Analysis and Applications* 38.2 (2017), pp. 649–655.


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.0
−0.5
0.0
0.5
1.0

Proof Strategy over $\mathbb{C}$

Characterize $\bigcup_{\substack{A\in\mathcal{P}_d^T \\ d\in\mathbb{Z}_{\geq 1}}} W(A)$: Let $S(\mathbb{C}^d)$ be the unit sphere. Then

$$\begin{aligned} & z \in \bigcup_{\substack{A\in\mathcal{P}_d^T \\ d\in\mathbb{Z}_{\geq 1}}} W(A) \\ \iff & \exists \vec{u} \in S(\mathbb{C}^d), P_1, P_2, \dots, P_T \in \mathcal{P}_d : \langle \vec{u}, P_T P_{T-1} \dots P_1 \vec{u} \rangle = z \\ \iff & \exists \vec{u}_0, \vec{u}_1, \dots, \vec{u}_T \in S(\mathbb{C}^d) : \langle \vec{u}_0, \vec{u}_T \rangle \langle \vec{u}_T, \vec{u}_{T-1} \rangle \dots \langle \vec{u}_1, \vec{u}_0 \rangle = z \end{aligned}$$

so the boundary of $\bigcup_{\substack{A\in\mathcal{P}_d^T \\ d\in\mathbb{Z}_{\geq 1}}} W(A)$ is given by critical points of the $\mathbb{R}$-smooth map $P : (S(\mathbb{C}^d))^T \to \mathbb{C}$ given by

$$(\vec{u}_0, \vec{u}_1, \dots, \vec{u}_T) \mapsto \langle \vec{u}_0, \vec{u}_T \rangle \langle \vec{u}_T, \vec{u}_{T-1} \rangle \dots \langle \vec{u}_1, \vec{u}_0 \rangle.$$

Extremizers must be coplanar, so enough to consider $d = 2$. Setting derivatives of each input to be parallel + algebra gives characterization of critical points (in terms of quaternions).

Proof strategy over $\mathbb{C}$: Relation to quaternions

Can rephrase problem and prove result in terms of quaternions.
To extremize

$$P(\vec{u}_0, \vec{u}_1, \ldots, \vec{u}_T) := \langle \vec{u}_0, \vec{u}_T \rangle \langle \vec{u}_T, \vec{u}_{T-1} \rangle \ldots \langle \vec{u}_1, \vec{u}_0 \rangle,$$

for unit quaternions $q_1, \ldots, q_T$, set $\vec{u}_t = q_t q_{t-1} \ldots q_1$ and let $q_0 = (q_T q_{T-1} \ldots q_1)^{-1}$. Then

$$P(\vec{u}_0, \ldots, \vec{u}_T) = \mathfrak{C} q_T \mathfrak{C} q_{T-1} \ldots \mathfrak{C} q_1 \mathfrak{C} q_0$$

where $\mathfrak{C}$ denotes complex part.
In other words, problem is to extremize $\mathfrak{C} q_T \mathfrak{C} q_{T-1} \ldots \mathfrak{C} q_1 \mathfrak{C} q_0$ subject to $q_T q_{T-1} \ldots q_1 q_0 = 1$.
Critical points (up to certain multiplication by complex units) are when (two of the q_t have zero complex part or):

- if $T+1$ is odd, $q_T = q_{T-1} = \cdots = q_1 = q_0$ is a $T+1$th quaternionic root of unity
- if $T+1$ is even, $q_T = q_{T-1} = \cdots = q_1 = q_0$ is a $2(T+1)$th quaternionic root of unity (where the multiplication by complex units is chosen to make their product 1, if necessary)

Proof Strategy over $\mathbb{R}$

For any $z \in \partial \left(\bigcup_{\substack{A \in \mathcal{P}_d^T \\ d \in \mathbb{Z}_{\geq 1}}} W(A) \right)$, Find a sequence of projections onto planes in $\mathbb{R}^4$ such that the extensions of these projections to $\mathbb{C}^4$ has an invariant copy of $\mathbb{C}^2$, and restricting there gives the complex projections realizing that z.

To do this, use the characterization of critical points to get a description of what these projections in $\mathbb{R}^4$ must look like.