Randomized Dimension Reduction with Statistical Guarantees

Yijun Dong
Oden Institute, University of Texas at Austin

Advisors: Prof. Per-Gunnar Martinsson, Prof. Rachel Ward

https://dyjdongyijun.github.io/defense.pdf

Roadmap

Computational efficiency

- Randomized pivoting-based interpolative & CUR decompositions
- Randomized subspace approximation: efficient canonical angle bounds & estimates

Sample efficiency

- Sample efficiency of data augmentation consistency regularization
- Adaptively weighted data augmentation consistency regularization for distributionally robust optimization under concept shift

Randomized Pivoting Algorithms for Interpolative and CUR Decompositions

Based on joint work with: Per-Gunnar Martinsson

Dong Y, Martinsson PG. Simpler is better: a comparative study of randomized algorithms for computing the CUR decomposition. arXiv preprint arXiv:2104.05877. 2021 Apr 13.

Matrix Skeleton Selection: Overview

- Inputs: $\mathbf{A} \in \mathbb{C}^{m \times n}$, target rank $k \leq \mathrm{rank}(\mathbf{A}) \Rightarrow$ Outputs: column/row skeletons $J_k \subseteq [n]$ and/or $I_k \subseteq [m]$

$$\mathbf{C} = \mathbf{A}(:, J_k) \quad \mathbf{R} = \mathbf{A}(I_k, :)$$

- Interpolative decomposition (ID)

$$\mathbf{A} \approx \mathbf{C}\left(\mathbf{C}\mathbf{A}^{\dagger}\right)$$

- CUR decomposition

$$\mathbf{A} \approx \mathbf{C}\left(\mathbf{C}^{\dagger}\mathbf{A}\mathbf{R}^{\dagger}\right)\mathbf{R}$$

Pivoting-based selection

- Column-pivoted QR
- (Strong) rank-revealing QR
- DEIM (SVD + LU with partial pivoting)

Sampling-based selection

- Uniform sampling
- Leverage score sampling
- Volume sampling

Key questions on **randomized pivoting-based skeleton selection**:

- Can we find **a general framework** that unifies the existing strategies?
- Are there more **efficient alternatives** to the existing algorithms?

Refer to (D., Martinsson, 2021) Section 3 for a brief survey

Randomized Pivoting-based Skeleton Selection: A General Framework

- Inputs: $\mathbf{A} \in \mathbb{C}^{m \times n}$, sample size l with $k < l \leq r = \mathrm{rank}(\mathbf{A})$, number of power iterations $q \in \{0,1,\cdots\}$
- Outputs: $J_l \subseteq [n]$ and/or $I_l \subseteq [m]$ such that $\mathbf{C} = \mathbf{A}(:,J_l) \in \mathbb{C}^{m \times l}$, $\mathbf{R} = \mathbf{A}(I_l,:) \in \mathbb{C}^{l \times n}$

Reduction stage: randomized dimension reduction / sketching

1. Draw randomized linear embedding $\mathbf{\Gamma} \sim P(\mathbb{C}^{l \times m})$ (e.g., $\Gamma_{ij} \sim \mathcal{N}(0,1/l)$ i.i.d.)
2. Construct a random row space approximate $\mathbf{X} \in \mathbb{C}^{l \times n}$
   1. Sketching on $\mathbf{A}$: $\mathbf{X} = \mathbf{\Gamma}\mathbf{A}(\mathbf{A}^*\mathbf{A})^q$ (with optional step-wise orthonormalization)
   2. Randomized SVD (RSVD) of $\mathbf{A}$: $[\sim,\sim,\mathbf{X}] = \mathrm{svd}(\mathbf{\Gamma}\mathbf{A}(\mathbf{A}^*\mathbf{A})^q)$ ←

- Gaussian matrices
- Subsampled randomized trigonometric transforms
- CountSketch
- Sparse sign matrices

- sketching is more efficient than RSVD by $O(nl^2)$

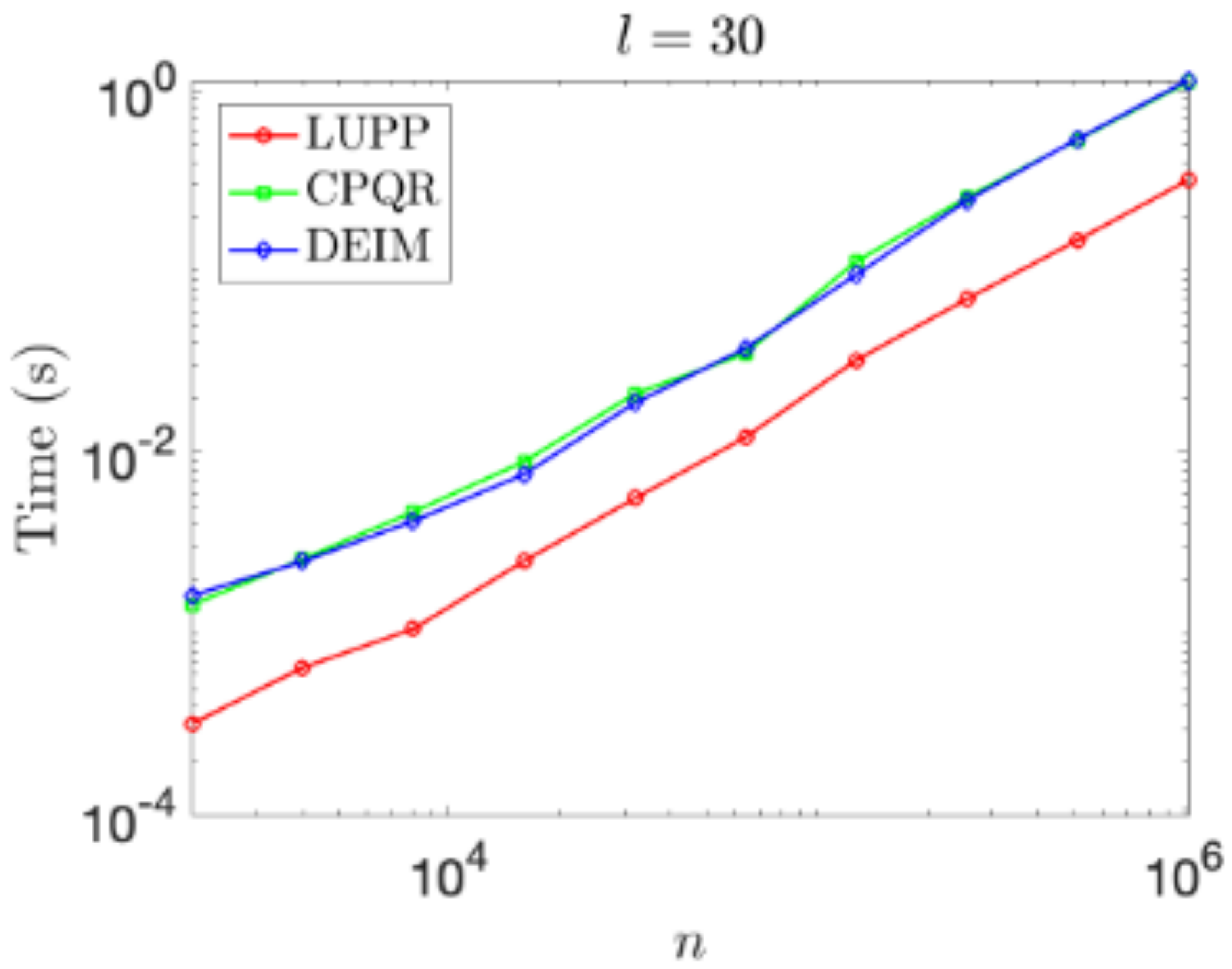
Pivoting stage: greedy skeleton selection

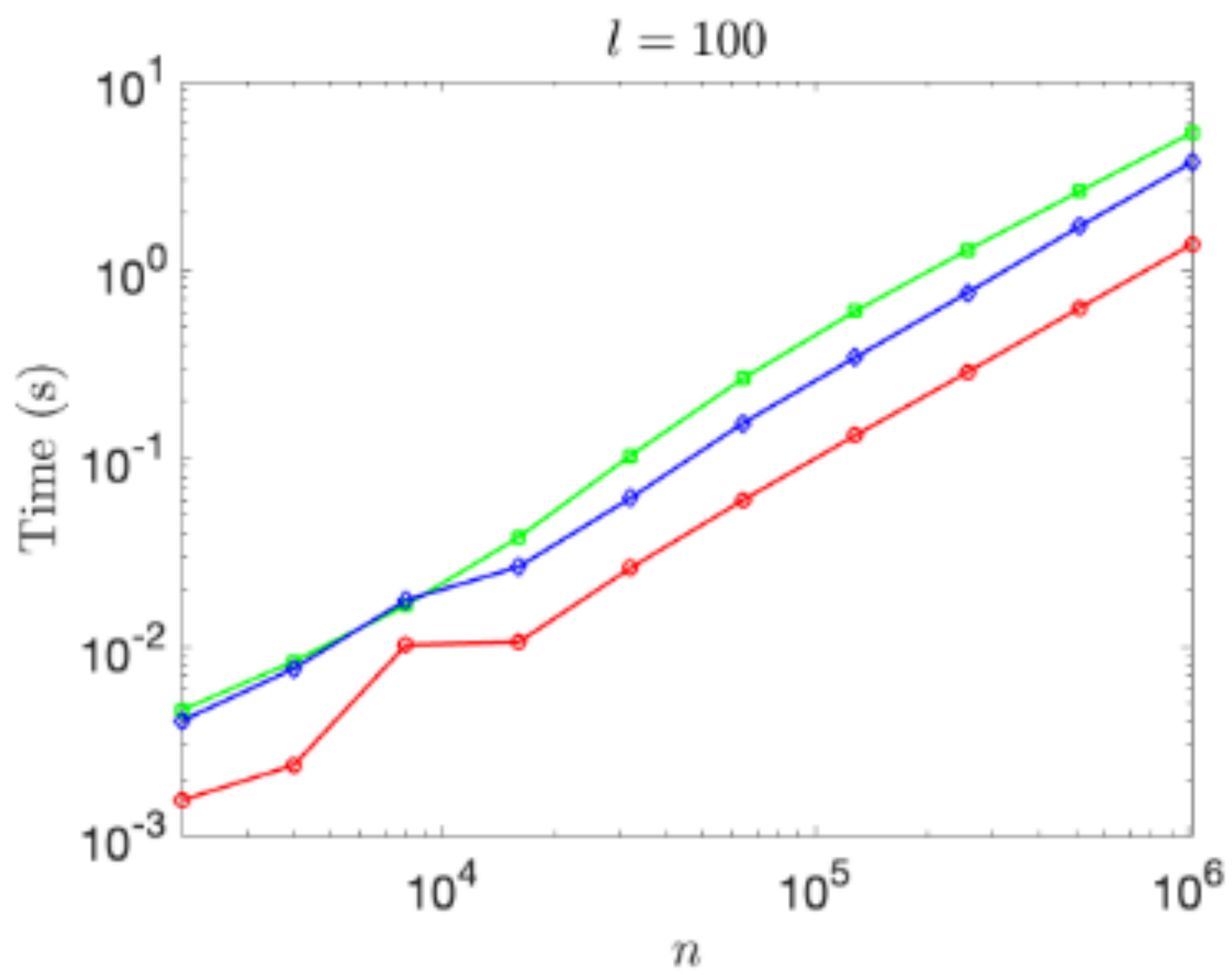
1. Column-wise pivoting on $\mathbf{X} \Rightarrow J_l$
2. (For CUR) row-wise pivoting on $\mathbf{C} = \mathbf{A}(:,J_l) \Rightarrow I_l$

Common pivoting schemes:

- Column pivoted QR (CPQR)
- LU with partial pivoting (LUPP)

Efficiency of Dimension Reduction + Pivoting



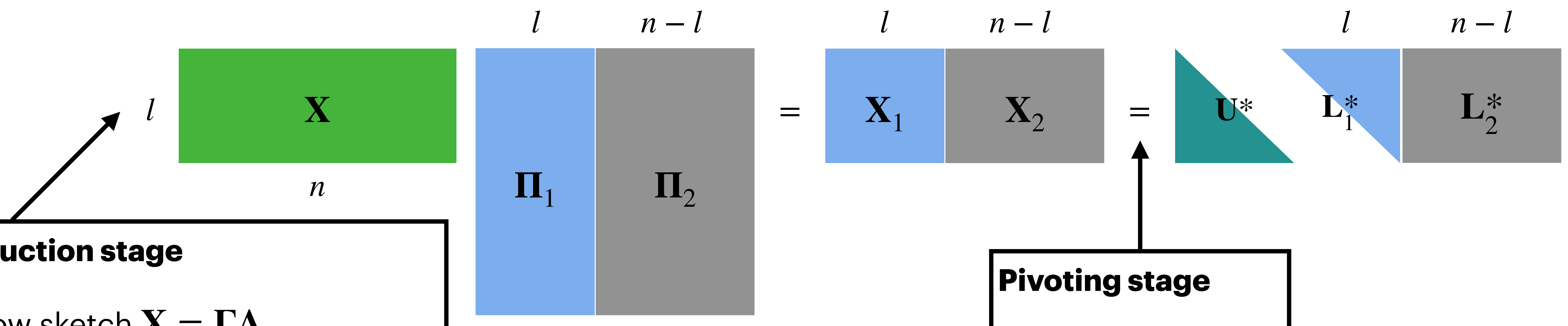


Runtime of sketching + LU with partial pivoting (LUPP) / sketching + column pivoted QR (CPQR) / randomized SVD + LUPP (DEIM) on $\mathbf{X} \in \mathbb{C}^{l \times n}$ column-wisely

- Runtime: **LUPP**≪DEIM<CPQR
- Despite the same asymptotic complexity $O(nl^2)$, LUPP is much **more efficient** than CPQR in practice
- LUPP and its variations (e.g., CALU) enjoy **better parallelizability** compared to CPQR

- LUPP is **less stable** than CPQR:
  - Not rank-revealing
  - Vulnerable to rank-deficiency
- **Randomization stabilizes LUPP!**
  - Sketching is “spectrum-revealing”
  - Sketching yields maximum-rank sample matrices almost-surely

Pivoting-based Skeletonization Error: Posterior Guarantee



Reduction stage

- Row sketch $\mathbf{X} = \mathbf{\Gamma A}$
- Approximated right singular vectors $[\sim, \sim, \mathbf{X}] = \mathrm{svd}(\mathbf{\Gamma A})$

Pivoting stage

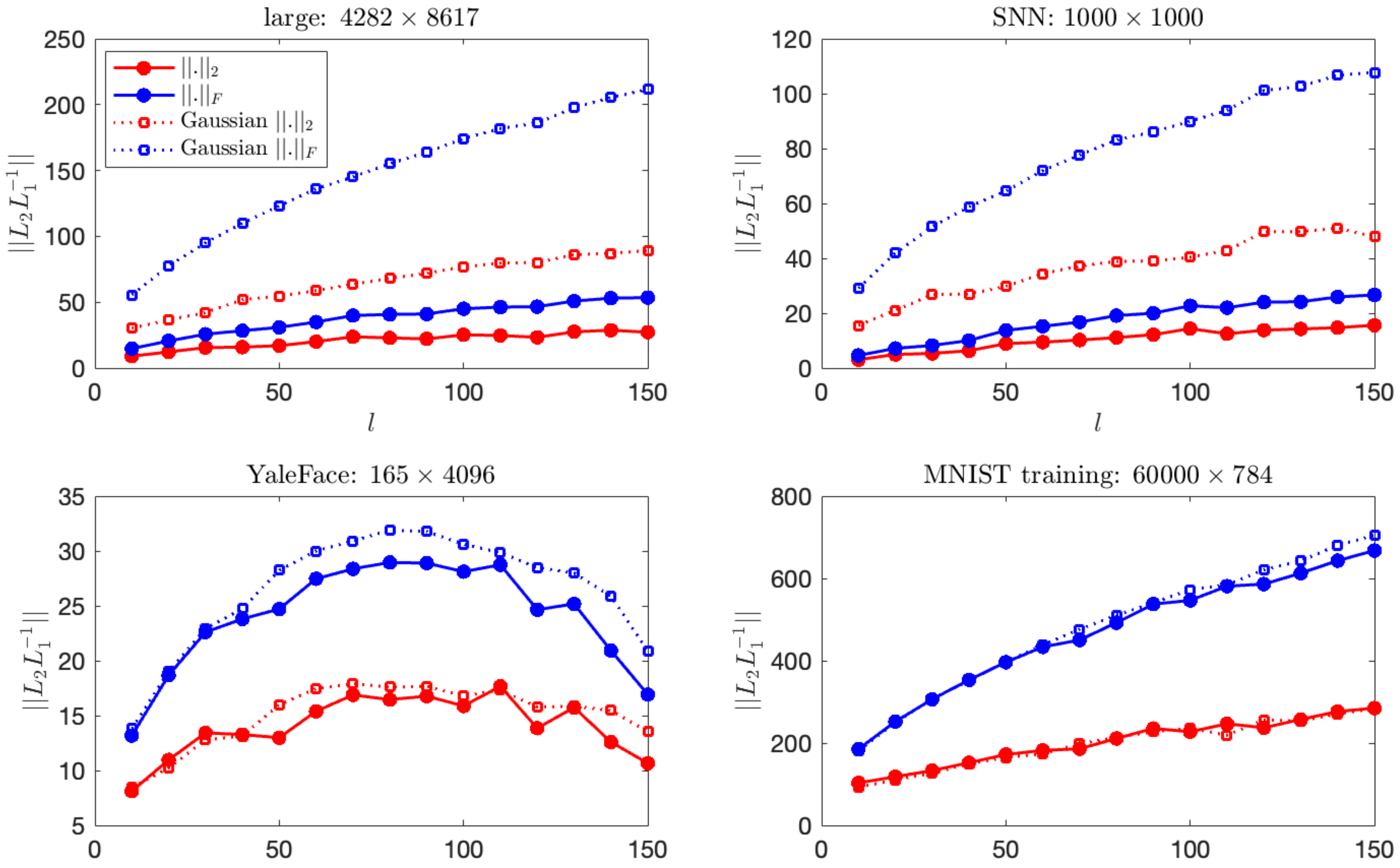
- LU-based: LUPP
- QR-based: CPQR

Theorem. (Posterior error guarantee of pivoting-based skeleton selection)

- Assume $\mathbf{X} \in \mathbb{C}^{l \times n}$ admits full row rank. Let $\mathbf{X}_1 \in \mathbb{C}^{l \times l}$ be the first l pivoted columns and $\mathbf{X}_2 \in \mathbb{C}^{l \times (n-l)}$ be the rest such that $\mathbf{X}\left[\mathbf{\Pi}_1, \mathbf{\Pi}_2\right] = \left[\mathbf{X}_1, \mathbf{X}_2\right]$. Then, for $\xi \in \{2, F\}$,

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\|_{\xi} \leq \underbrace{\eta}_{\text{pivoting error}} \cdot \underbrace{\|\mathbf{A} - \mathbf{A}\mathbf{X}^{\dagger}\mathbf{X}\|_{\xi}}_{\text{reduction error}}, \quad \eta \leq \sqrt{1 + \|\mathbf{X}_1^{\dagger}\mathbf{X}_2\|_2^2}$$

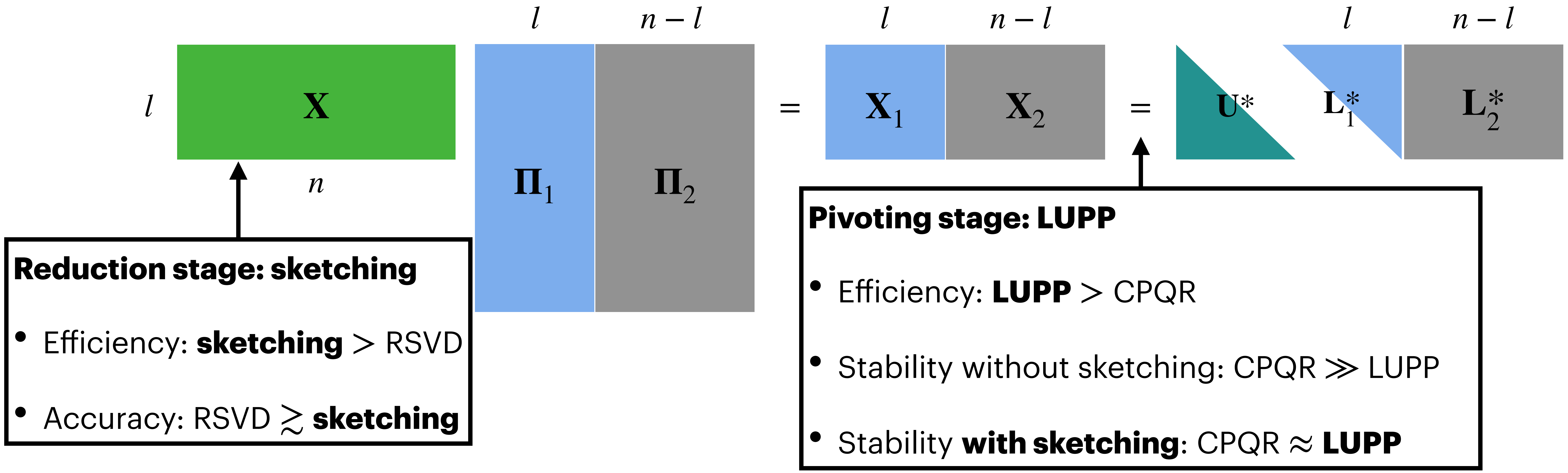
Randomization Stabilizes LUPP: $\eta = O(l)$ in practice



- **Worst-case** pivoting error factor: $\eta = \Theta(2^l)$ for both LUPP and CPQR (e.g., Kahan matrix)
- With randomization via sketching, we observe $\eta = O(l)$ **in practice** (cf. Trefethen et al, 1990)

Trefethen LN, Schreiber RS. Average-case stability of Gaussian elimination. SIAM Journal on Matrix Analysis and Applications. 1990 Jul;11(3):335-60.

Efficient Alternative: Sketching + LUPP



Existing algorithms

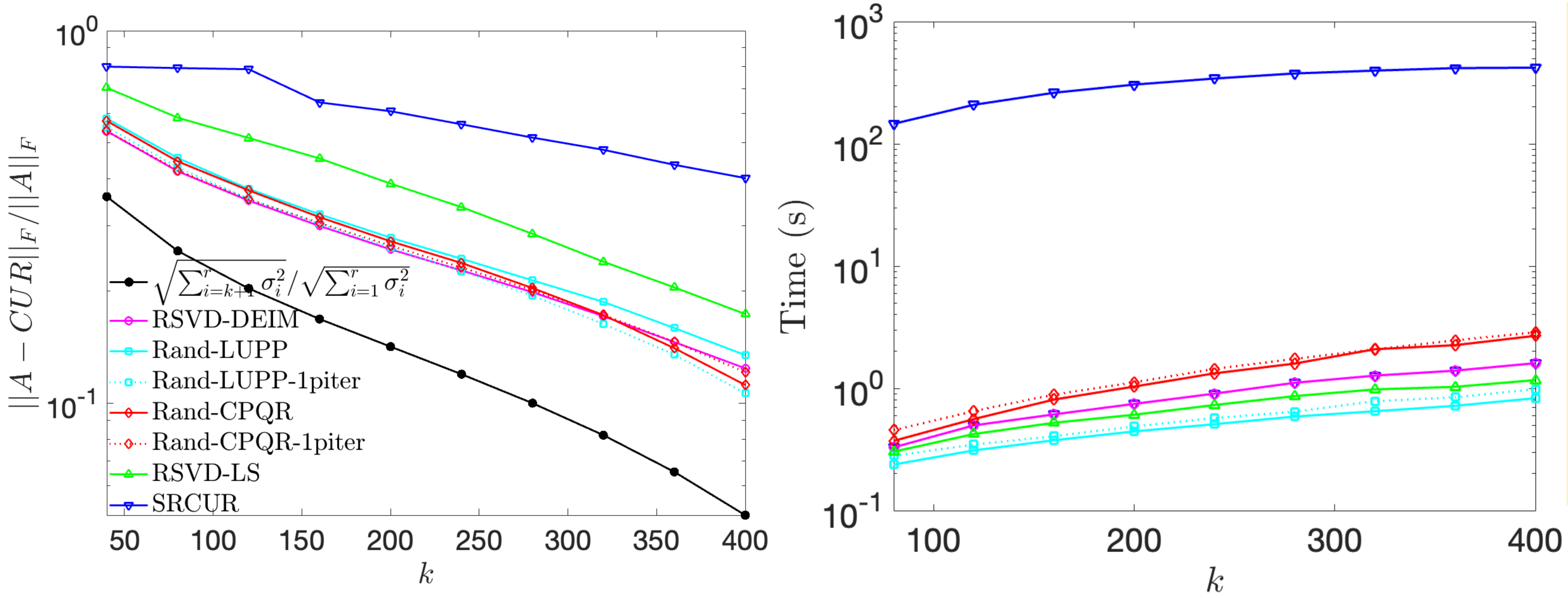
- (Voronin et al, 2017): Sketching + CPQR
- (Sorensen et al, 2016): (R)SVD + LUPP (DEIM)

More efficient alternative: **Sketching + LUPP**

- Sketching stage: $\mathbf{X} = \mathbf{\Gamma}\mathbf{A}$
- Pivoting stage: LUPP on $\mathbf{X}$ column-wisely

1. Voronin S, Martinsson PG. Efficient algorithms for CUR and interpolative matrix decompositions. Advances in Computational Mathematics. 2017 Jun;43:495-516.
2. Sorensen DC, Embree M. A deim induced cur factorization. SIAM Journal on Scientific Computing. 2016;38(3):A1454-82.

Accuracy & Efficiency of Sketching + LUPP: MNIST



Accuracy

- **Rand-LUPP** ≈ RSVD-DEIM ≈ Rand-CPQR > RSVD-LS > SRCUR
- $q = 1$ boosts accuracy sufficiently

Efficiency

- **Rand-LUPP** > RSVD-LS > RSVD-DEIM > Rand-CPQR > SRCUR

- Rand-LUPP (ours): sketching + LUPP (with $q = 1$ power iteration)
- Rand-CPQR: sketching + CPQR (with $q = 1$ power iteration)
- RSVD-DEIM: RSVD + LUPP
- RSVD-LS: leverage score with approximated singular vectors from RSVD
- SRCUR: spectrum-revealing CUR (Chen et al, 2020) based on spectrum-revealing pivoting schemes

Chen C, Gu M, Zhang Z, Zhang W, Yu Y. Efficient spectrum-revealing CUR matrix decomposition. InInternational Conference on Artificial Intelligence and Statistics 2020 Jun 3 (pp. 766-775). PMLR.

Randomized Subspace Approximations: Efficient Bounds and Estimates for Canonical Angles

Based on joint work with: Per-Gunnar Martinsson, Yuji Nakatsukasa

Dong Y, Martinsson PG, Nakatsukasa Y. Efficient Bounds and Estimates for Canonical Angles in Randomized Subspace Approximations. arXiv preprint arXiv:2211.04676. 2022 Nov 9.

Leading Singular Subspaces

- Singular value decomposition (SVD)

  Given $\mathbf{A} \in \mathbb{C}^{m \times n}$, $1 \leq k \leq r = \mathrm{rank}(\mathbf{A})$, rank-k truncated SVD:

$$\mathbf{A}_k = \underset{m\times k}{\mathbf{U}_k} \; \underset{k\times k}{\mathbf{\Sigma}_k} \; \underset{k\times n}{\mathbf{V}_k^*}$$

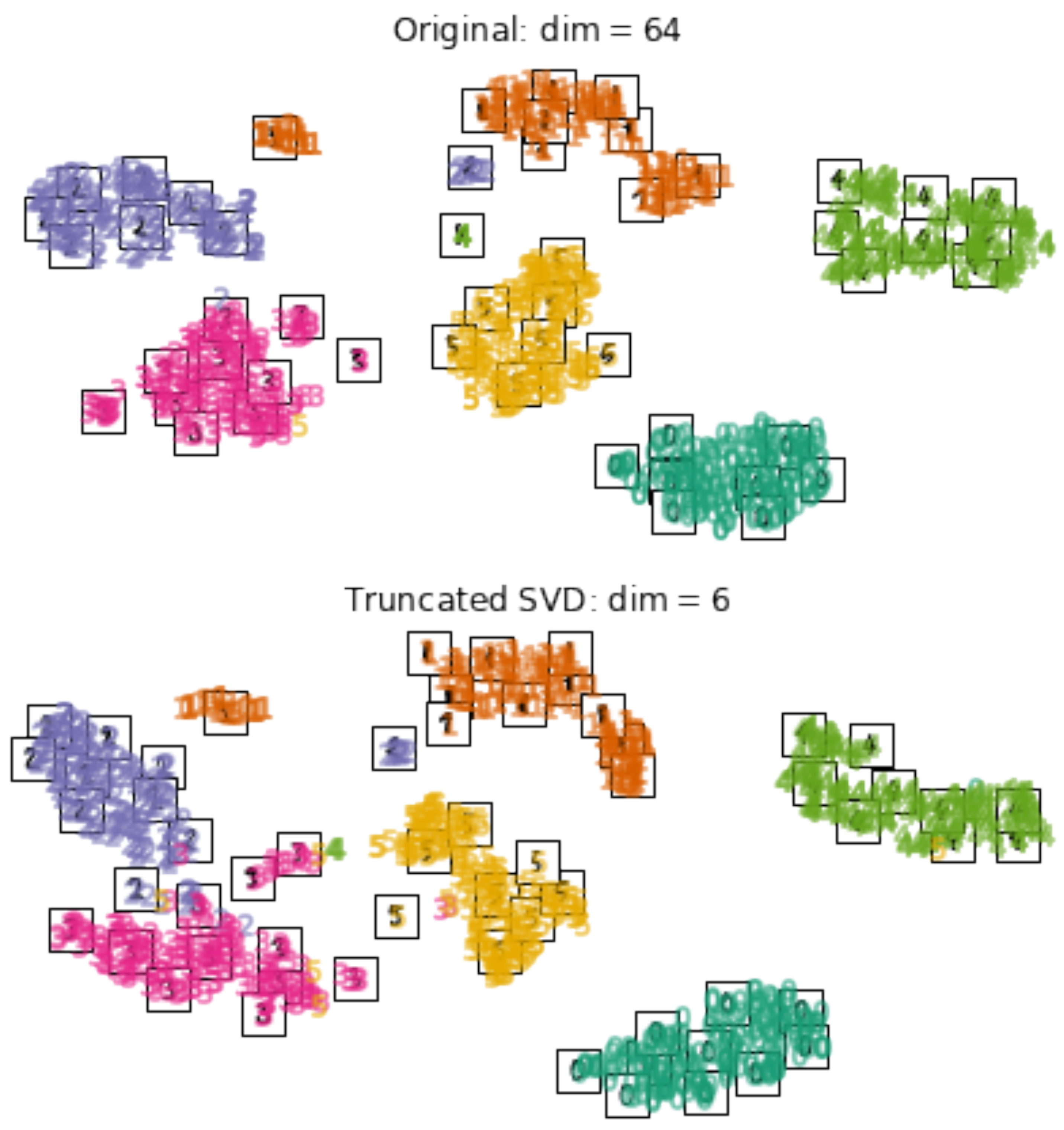
  - Maximum-k singular values: $\mathbf{\Sigma}_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$
  - **Leading-k singular subspaces**: $\mathbf{U}_k^* \mathbf{U}_k = \mathbf{V}_k^* \mathbf{V}_k = \mathbf{I}_k$

- Eckart–Young–Mirsky theorem

$$\mathbf{A}_k = \min_{\mathrm{rank}(\widehat{\mathbf{A}}) \leq k} \|\mathbf{A} - \widehat{\mathbf{A}}\|_F$$

  - Truncated SVD provides the optimal rank-k approximation

- Broad Applications
  - Low-rank approximations, PCA, CCA, spectral clustering, leverage score sampling, etc.



Spectral clustering on the dimension-6 leading singular subspace of a mini-MNIST dataset (8 × 8 images of digits 0-5)

Sketching: Approximate leading singular subspaces efficiently for large matrices

Questions: How accurate are these approximations? Tight & efficiently computable error bounds & estimates?

Randomized Subspace Approximations with Sketching

- Inputs: $\mathbf{A} \in \mathbb{C}^{m \times n}$, sample size l with $k < l \leq r = \mathrm{rank}(\mathbf{A})$ (e.g., $l = 2k \ll r$), number of power iterations $q \in \{0,1,2,\cdots\}$ ($q \leq 2$ usually)
- Outputs: $\mathrm{RSVD}(\mathbf{A}, l, q) = (\widehat{\mathbf{U}}_l \in \mathbb{C}^{m\times l}, \widehat{\mathbf{\Sigma}}_l \in \mathbb{C}^{l\times l}, \widehat{\mathbf{V}}_l \in \mathbb{C}^{n\times l})$ such that $\widehat{\mathbf{A}}_l = \widehat{\mathbf{U}}_l \widehat{\mathbf{\Sigma}}_l \widehat{\mathbf{V}}_l^* \approx \mathbf{A}$

1. **Randomized linear embedding** (Johnson-Lindenstrauss transforms, etc.)
   - Draw $\mathbf{\Omega} \sim P(\mathbb{C}^{n\times l})$ with i.i.d. entries $\Omega_{ij} \sim \mathcal{N}(0, l^{-1})$ such that $\mathbb{E}[\mathbf{\Omega}\mathbf{\Omega}^*] = \mathbf{I}_n$

   Isotropic embedding: uniform over unit sphere

2. **Sketching** with power iterations
   - Randomized **power** iterations (unstable): $\mathbf{X}^{(q)} = (\mathbf{A}\mathbf{A}^*)^q \mathbf{A}\mathbf{\Omega}$
   - Randomized **subspace** iterations (stable): $\mathbf{X}^{(0)} = \mathrm{ortho}(\mathbf{A}\mathbf{\Omega})$, $\mathbf{X}^{(i)} = \mathrm{ortho}(\mathbf{A}\ \mathrm{ortho}(\mathbf{A}^*\mathbf{X}^{(i-1)}))\ \forall\ i \in [q]$
3. $\mathbf{Q}_X = \mathrm{ortho}(\mathbf{X}^{(q)})$
4. $[\widetilde{\mathbf{U}}_l, \widehat{\mathbf{\Sigma}}_l, \widehat{\mathbf{V}}_l] = \mathrm{svd}(\mathbf{A}^*\mathbf{Q}_X)$
5. $\widehat{\mathbf{U}}_l = \mathbf{Q}_X \widetilde{\mathbf{U}}_l$

Key observations: with $\mathbf{\Sigma}$ being the spectrum of $\mathbf{A}$

- For any $q \in \mathbb{N}$, q power iterations correspond to $\mathbf{\Sigma}^{2q+1}$
- Compared to $\widehat{\mathbf{U}}_l$, $\widehat{\mathbf{V}}_l$ enjoys half more power iterations (i.e., $\mathbf{\Sigma}^{2q+2}$)

Canonical Angles: Alignment between Subspaces

- Canonical angles $\angle(\mathscr{U}, \mathscr{V}) = (\theta_1, \cdots, \theta_k)$ measure the alignment between two subspaces $\mathscr{U}, \mathscr{V} \subseteq \mathbb{C}^d$ with dimensions $k, l \leq d$ respectively ($k < l$ w.l.o.g), e.g.,
  - True leading singular subspace: $\mathscr{U} = \mathrm{range}(\mathbf{U}_k)$
  - Approximated leading singular subspace: $\mathscr{V} = \mathrm{range}(\widehat{\mathbf{U}}_l)$
- Left & right **canonical angles** of $\mathrm{RSVD}(\mathbf{A}, l, q) = (\widehat{\mathbf{U}}_l, \widehat{\boldsymbol{\Sigma}}_l, \widehat{\mathbf{V}}_l)$: $\forall\ i \in [k]$,

$$\sin\angle_i(\mathbf{U}_k, \widehat{\mathbf{U}}_l) = \sigma_{k-i+1}((\mathbf{I}_m - \widehat{\mathbf{U}}_l\widehat{\mathbf{U}}_l^*)\mathbf{U}_k), \quad \cos\angle_i(\mathbf{U}_k, \widehat{\mathbf{U}}_l) = \sigma_i(\widehat{\mathbf{U}}_l^*\mathbf{U}_k)$$

$$\sin\angle_i(\mathbf{V}_k, \widehat{\mathbf{V}}_l) = \sigma_{k-i+1}((\mathbf{I}_m - \widehat{\mathbf{V}}_l\widehat{\mathbf{V}}_l^*)\mathbf{V}_k), \quad \cos\angle_i(\mathbf{V}_k, \widehat{\mathbf{V}}_l) = \sigma_i(\widehat{\mathbf{V}}_l^*\mathbf{V}_k)$$

Prior v.s. **posterior** guarantees: computed without v.s. with the outputs $(\widehat{\mathbf{U}}_l, \widehat{\boldsymbol{\Sigma}}_l, \widehat{\mathbf{V}}_l)$

- Prior guarantees are probabilistic, with randomness from $\boldsymbol{\Omega} \sim P(\mathbb{C}^{n\times l})$
- Posterior guarantees are deterministic with given $(\widehat{\mathbf{U}}_l, \widehat{\boldsymbol{\Sigma}}_l, \widehat{\mathbf{V}}_l)$

Space-agnostic Prior Probabilistic Bounds

Theorem. (Space-agnostic bounds under multiplicative oversampling. (D., Martinsson, Nakatsukasa, 2022))

- With Gaussian embedding; small $q \in \mathbb{N}$ such that $\eta \triangleq \left(\sum_{j=k+1}^{r} \sigma_j^{4q+4}\right)^2 \Big/ \sum_{j=k+1}^{r} \sigma_j^{2(4q+4)} = \Omega(l)$; oversampling $l = \Omega(k)$
  - Notice that $1 < \eta \leq r - k$ and usually $r - k \gg l$. $\eta = \Omega(l)$ refers to a realistic case with non-negligible approximation error: when the tail of the spectrum $\{\sigma_j\}_{j=k+1}^r$ remains non-trivial after q power iterations
- With high probability (at least $1 - e^{-\Theta(k)} - e^{-\Theta(l)}$), there exist $\epsilon_1 = \Theta(\sqrt{k/l})$, $\epsilon_2 = \Theta(\sqrt{l/\eta})$, $\epsilon_1, \epsilon_2 \in (0,1)$ such that, $\forall\ i \in [k]$

$$\left(1 + O_{\epsilon_1,\epsilon_2}\left(\frac{l \cdot \sigma_i^{4q+2}}{\sum_{j=k+1}^r \sigma_j^{4q+2}}\right)\right)^{-\frac{1}{2}} \leq \sin\angle_i(\mathbf{U}_k, \widehat{\mathbf{U}}_l) \leq \left(1 + \frac{1-\epsilon_1}{1+\epsilon_2} \cdot \frac{l \cdot \sigma_i^{4q+2}}{\sum_{j=k+1}^r \sigma_j^{4q+2}}\right)^{-\frac{1}{2}}$$

$$\left(1 + O_{\epsilon_1,\epsilon_2}\left(\frac{l \cdot \sigma_i^{4q+4}}{\sum_{j=k+1}^r \sigma_j^{4q+4}}\right)\right)^{-\frac{1}{2}} \leq \sin\angle_i(\mathbf{V}_k, \widehat{\mathbf{V}}_l) \leq \left(1 + \frac{1-\epsilon_1}{1+\epsilon_2} \cdot \frac{l \cdot \sigma_i^{4q+4}}{\sum_{j=k+1}^r \sigma_j^{4q+4}}\right)^{-\frac{1}{2}}$$

- In practice, taking $\epsilon_1 = \sqrt{k/l}$, $\epsilon_2 = \sqrt{l/(r-k)}$ is sufficient for upper bounds when $l \geq 1.6k$ and $q \leq 10$

Comparison with Existing Prior Probabilistic Guarantees

- Given $\boldsymbol{\Omega} \sim P(\mathbb{C}^{n\times l})$, let $\boldsymbol{\Omega}_1 \triangleq \mathbf{V}_k^* \boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_2 \triangleq \mathbf{V}_{r\backslash k}^* \boldsymbol{\Omega}$. Then, $\boldsymbol{\Omega}_1 \sim P(\mathbb{C}^{k\times l})$ and $\boldsymbol{\Omega}_2 \sim P(\mathbb{C}^{(r-k)\times l})$
- **Prior work (Saibaba, 2018)**[1]:

$$\sin\angle_i(\mathbf{U}_k, \widehat{\mathbf{U}}_l) \leq \left(1 + \frac{\sigma_i^{4q+2}}{\sigma_{k+1}^{4q+2}\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2^2}\right)^{-\frac{1}{2}}, \quad \sin\angle_i(\mathbf{V}_k, \widehat{\mathbf{V}}_l) \leq \left(1 + \frac{\sigma_i^{4q+4}}{\sigma_{k+1}^{4q+4}\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2^2}\right)^{-\frac{1}{2}}$$

where for $l \geq k+2$, given any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2 \leq \frac{e\sqrt{l}}{l-k+1}\left(\frac{2}{\delta}\right)^{\frac{1}{l-k+1}}\left(\sqrt{n-k} + \sqrt{l} + \sqrt{2\log\frac{2}{\delta}}\right) = \Omega\left(\sqrt{\frac{n-k}{l}}\right)$$

Recall the correspondence in Theorem 1:

$$\frac{1}{l}\sum_{j=k+1}^{r}\sigma_j^{4q+2} \leq \frac{n-k}{l}\sigma_{k+1}^{4q+2}$$

where the smaller values lead to the tighter upper bounds

- Theorem 1 is **space-agnostic** since the randomized linear embedding $\boldsymbol{\Omega} \sim P(\mathbb{C}^{n\times l})$ is **isotropic**
  - Only depends on the spectrum $\{\sigma_j\}_{j=1}^r$, but not on the singular subspaces $(\mathbf{U}_k, \mathbf{U}_{r\backslash k})$ or $(\mathbf{V}_k, \mathbf{V}_{r\backslash k})$
  - In proof, we took an integrated view on the concentration of $\boldsymbol{\Sigma}_{r\backslash k}^{2q+1}\boldsymbol{\Omega}_2$

Saibaba, Arvind K. "Randomized subspace iteration: Analysis of canonical angles and unitarily invariant norms." *SIAM Journal on Matrix Analysis and Applications* 40.1 (2019): 23-48.

Unbiased Space-agnostic Estimates

- Draw independent Gaussian random matrices $\left\{ \boldsymbol{\Omega}_1^{(j)} \sim P(\mathbb{C}^{k \times l}) \,\middle|\, j \in [N] \right\}$ and $\left\{ \boldsymbol{\Omega}_2^{(j)} \sim P(\mathbb{C}^{(r-k) \times l}) \,\middle|\, j \in [N] \right\}$
- Unbiased canonical angle estimates $\alpha_i = \mathbb{E}\left[\sin\angle_i(\mathbf{U}_k, \widehat{\mathbf{U}}_l)\right]$, $\beta_i = \mathbb{E}\left[\sin\angle_i(\mathbf{V}_k, \widehat{\mathbf{V}}_l)\right]$ $\forall\ i \in [k]$ such that

$$\sin\angle_i(\mathbf{U}_k, \widehat{\mathbf{U}}_l) \approx \alpha_i = \frac{1}{N}\sum_{j=1}^{N}\left(1 + \sigma_i^2 \left(\boldsymbol{\Sigma}_k^{2q+1}\boldsymbol{\Omega}_1^{(j)}\left(\boldsymbol{\Sigma}_{r\backslash k}^{2q+1}\boldsymbol{\Omega}_2^{(j)}\right)^{\dagger}\right)\right)^{-\frac{1}{2}}$$

Corresponds to $\dfrac{1 \mp \epsilon_1}{1 \pm \epsilon_2} \cdot \dfrac{l \cdot \sigma_i^{4q+2}}{\sum_{j=k+1}^{r} \sigma_j^{4q+2}}$ in the upper/lower bounds of Theorem 1

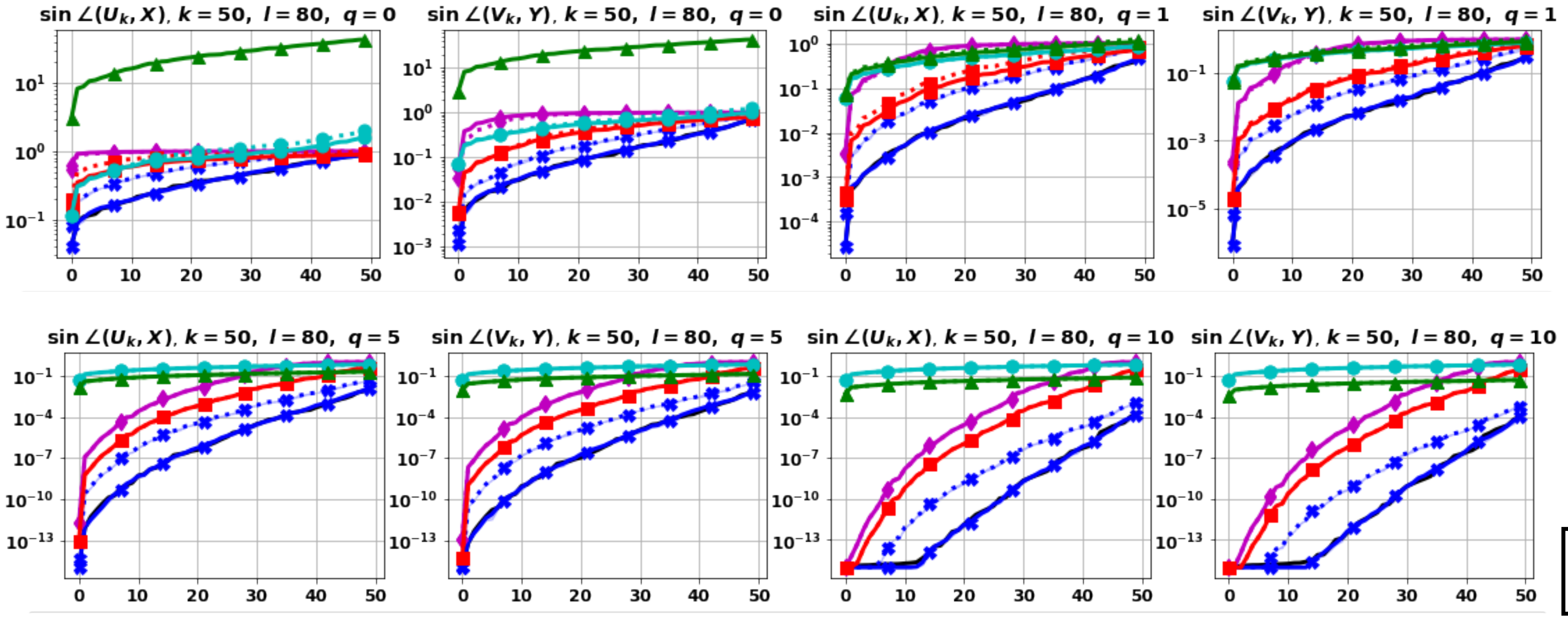
$$\sin\angle_i(\mathbf{V}_k, \widehat{\mathbf{V}}_l) \approx \beta_i = \frac{1}{N}\sum_{j=1}^{N}\left(1 + \sigma_i^2 \left(\boldsymbol{\Sigma}_k^{2q+2}\boldsymbol{\Omega}_1^{(j)}\left(\boldsymbol{\Sigma}_{r\backslash k}^{2q+2}\boldsymbol{\Omega}_2^{(j)}\right)^{\dagger}\right)\right)^{-\frac{1}{2}}$$

- **Low variance** in practice (i.e., negligible when $N \geq 3$)
- **Can be computed efficiently** with $O(rl^2)$ operations (for a given spectrum $\boldsymbol{\Sigma}$)
- **For any** $k \leq l \leq r$, without further assumptions on the sample size (e.g., $\eta = \Omega(l), l = \Omega(k)$)

Posterior Residual-based Guarantees

1. Posterior bounds based on full residuals: Theorem 2. (D., Martinsson, Nakatsukasa, 2022)
   - $\sin\angle_i(\mathbf{U}_k, \widehat{\mathbf{U}}_l) \leq \dfrac{\sigma_{k-i+1}\left(\left(\mathbf{I}_m - \widehat{\mathbf{U}}_l \widehat{\mathbf{U}}_l^*\right)\mathbf{A}\right)}{\sigma_k} \wedge \dfrac{\sigma_1\left(\left(\mathbf{I}_m - \widehat{\mathbf{U}}_l \widehat{\mathbf{U}}_l^*\right)\mathbf{A}\right)}{\sigma_i}$
   - Deterministic and **algorithm-independent** (e.g., holds for any $k \leq l \leq r$, and any embedding $\mathbf{\Omega}$)
   - Can be approximated with $O(mnl)$ operations
2. Posterior bounds based on sub-residuals: Theorem 3.
   - Let $\mathbf{E}_{31} \triangleq \widehat{\mathbf{U}}^*_{m\backslash l}\mathbf{A}\widehat{\mathbf{V}}_k$, $\mathbf{E}_{32} \triangleq \widehat{\mathbf{U}}^*_{m\backslash l}\mathbf{A}\widehat{\mathbf{V}}_{l\backslash k}$, $\mathbf{E}_{33} \triangleq \widehat{\mathbf{U}}^*_{m\backslash l}\mathbf{A}\widehat{\mathbf{V}}_{n\backslash l}$, $\Gamma_1 \triangleq \dfrac{\sigma_k^2 - \|\mathbf{E}_{33}\|_2^2}{\sigma_k}$, $\Gamma_2 \triangleq \dfrac{\sigma_k^2 - \|\mathbf{E}_{33}\|_2^2}{\|\mathbf{E}_{33}\|_2}$.
     Assume $\sigma_k > \|\mathbf{E}_{33}\|_2$. Then, for any unitary invariant norm $\|\cdot\|$, $\|\sin\angle(\mathbf{U}_k, \widehat{\mathbf{U}}_l)\| \leq \|[\mathbf{E}_{31}, \mathbf{E}_{32}]\| \big/ \Gamma_1$
   - Deterministic and holds for any $k \leq l \leq r$, and any embedding $\mathbf{\Omega}$
   - Can be approximated with $O(mnl)$ operations

Space-agnostic bounds & estimates win on MNIST: Polynomial spectral decay



$l = 1.6k$

$q \in \{0,1\}$

$l = 1.6k$

$q \in \{5,10\}$

Black line: true canonical angles

Blue lines/dashes (with shade): unbiased space-agnostic estimates computed with true/approximated singular values

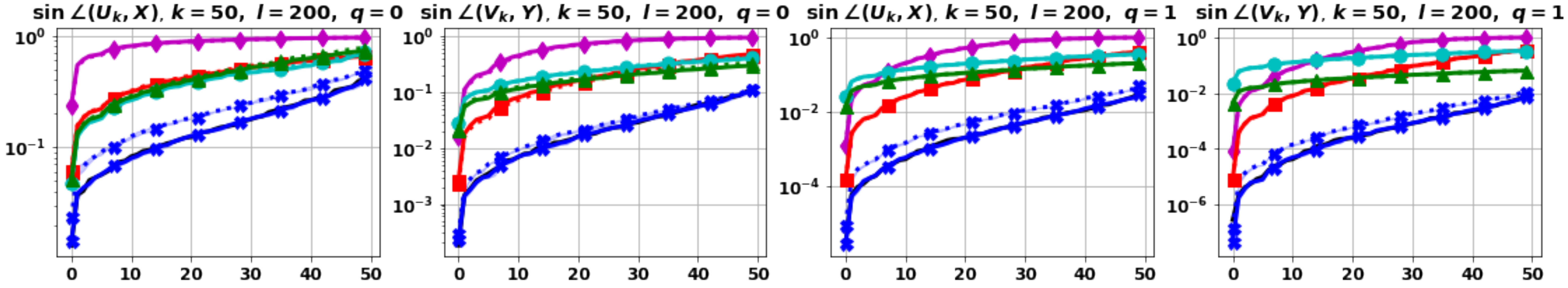
Red lines/dashes: space-agnostic upper bounds with true/approximated singular values, $\epsilon_1 = \sqrt{k/l}, \epsilon_2 = \sqrt{l/(r-k)}$

Magenta lines/dashes: (Saibaba, 2018) bounds with true/approximated singular values and the true singular subspaces

Cyan & green lines/dashes: Posterior residual-based bounds in Theorem 2 & 3 with true/approximated singular values

shade = min/max in $N = 3$ samples ⇒ negligible variance!

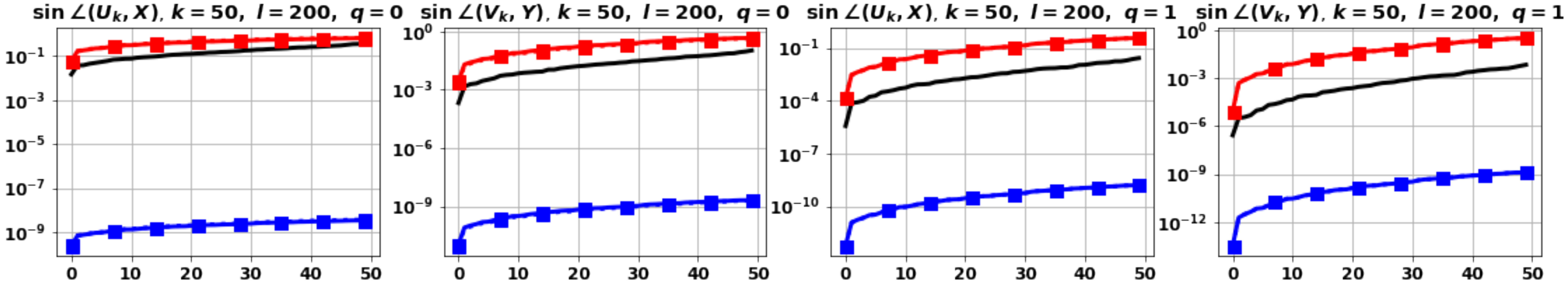
How about space-agnostic lower bounds in practice: MNIST



$l = 4k$

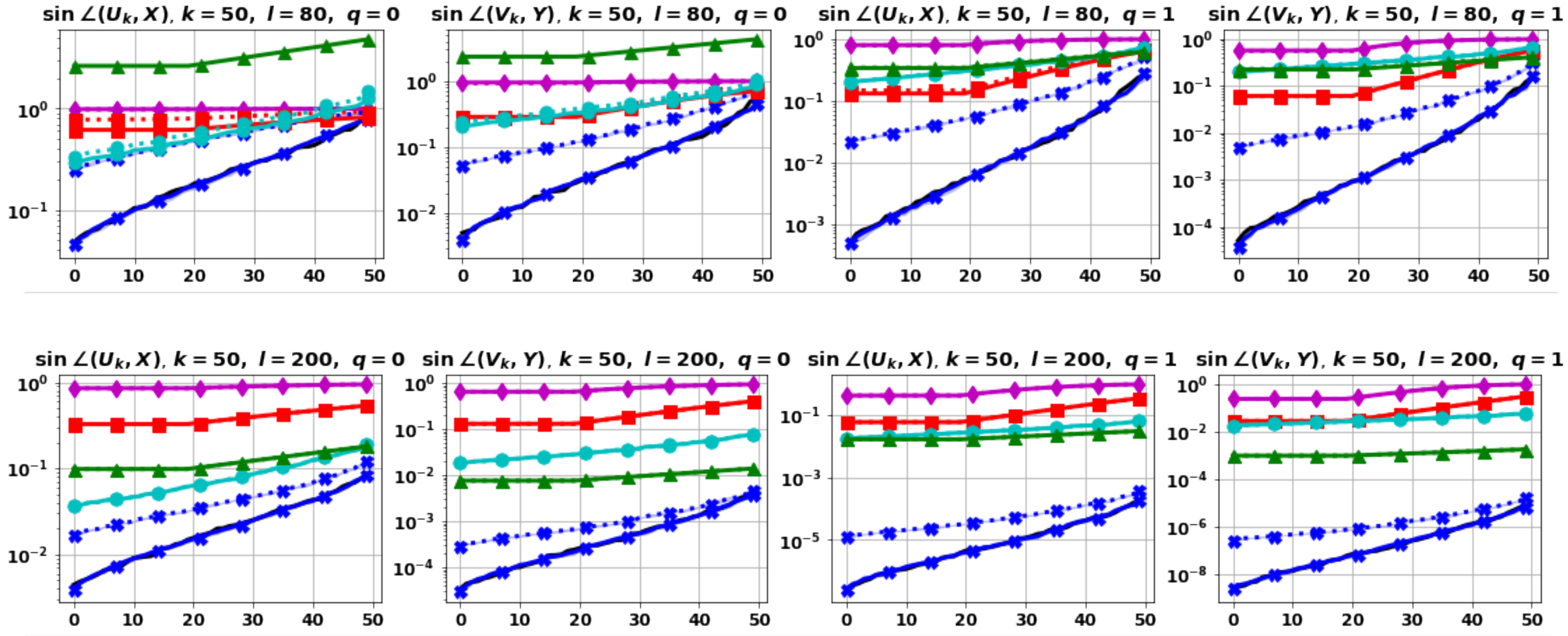
$q \in \{0,1\}$

Unbiased space-agnostic estimates, space-agnostic upper bounds, (Saibaba, 2018) bounds, Posterior residual-based bounds in Theorem 2 & 3 (with true/approximated singular values), and true canonical angles



Space-agnostic upper bounds and lower bounds with true singular values and $\epsilon_1 = \sqrt{k/l}, \epsilon_2 = \sqrt{l/(r-k)}$

When are posterior bounds more effective: Exponential spectral decay + low-error regimes



$l = 1.6k$

$q \in \{0,1\}$

$l = 4k$

$q \in \{0,1\}$

Unbiased space-agnostic estimates, space-agnostic upper bounds, (Saibaba, 2018) bounds, Posterior residual-based bounds in Theorem 2 & 3 (with true/approximated singular values), and true canonical angles

Sample Efficiency of Data Augmentation Consistency Regularization

Based on joint work with: Shuo Yang, Rachel Ward, Inderjit Dhillon, Sujay Sanghavi, Qi Lei

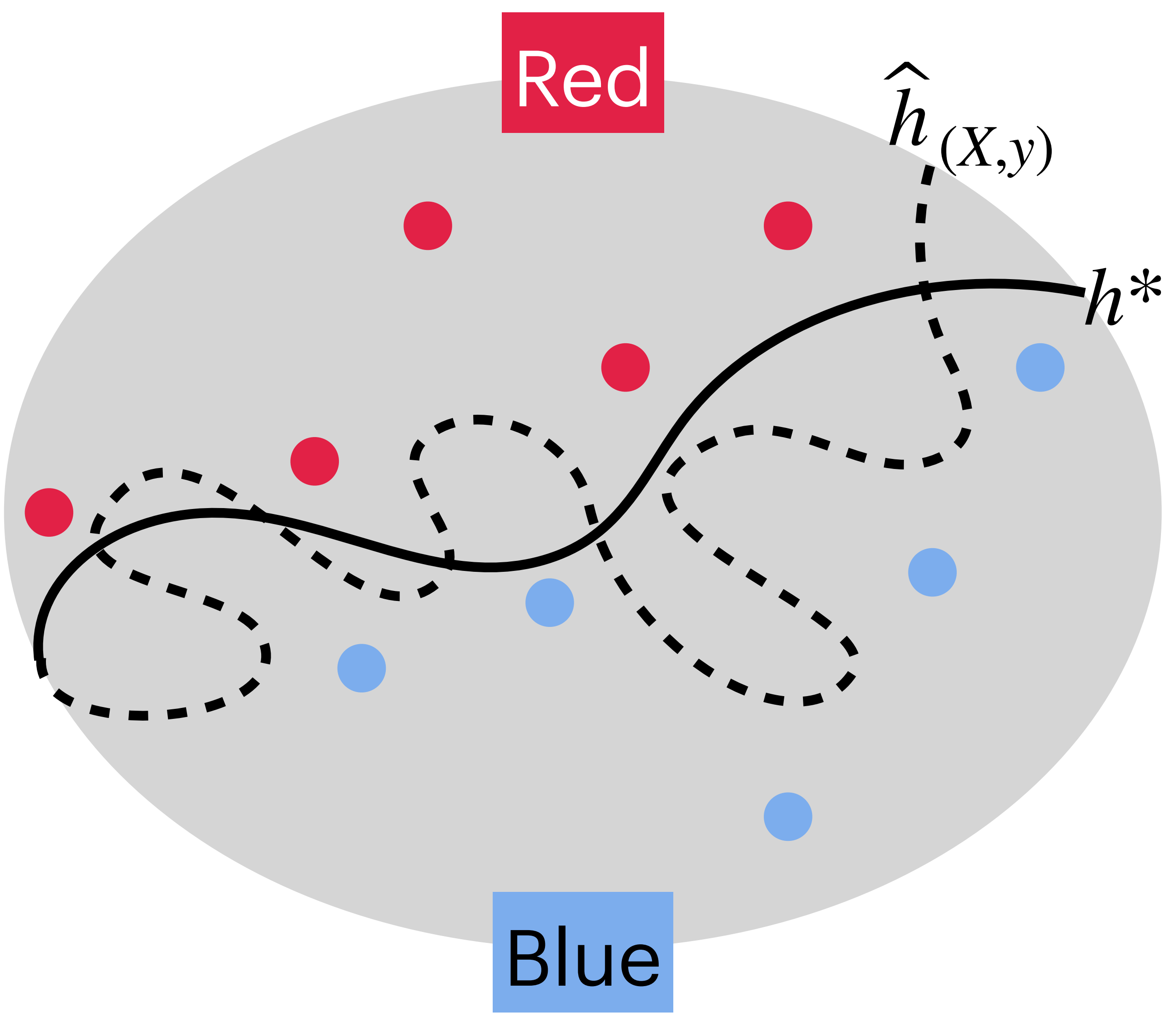
Yang S, Dong Y, Ward R, Dhillon IS, Sanghavi S, Lei Q. Sample efficiency of data augmentation consistency regularization. arXiv preprint arXiv:2202.12230. 2022 Feb 24.

Generalization & Sample Complexity

Learn **unknown population**

- As ground truth distribution $P : \mathscr{X} \times \mathscr{Y} \to [0,1]$
- Within a function (hypothesis) class $\mathscr{H} \ni h : \mathscr{X} \to \mathscr{Y}$
- Through a proper loss function $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$ with ground truth $h^* \triangleq \underset{h \in \mathscr{H}}{\operatorname{argmin}} \left\{ L(h) \triangleq \mathbb{E}_{(x,y)\sim P} \left[\ell(h(x), y) \right] \right\}$

Population risk



From **limited samples**

- As training data $(X, y) = \left\{ (x_i, y_i) \right\}_{i=1}^{n} \sim P(x, y)^n$
- Via learning algorithm $\mathscr{L}$, e.g., **empirical risk minimization**

(ERM): $\hat{h}_{(X,y)} \triangleq \underset{h \in \mathscr{H}}{\operatorname{argmin}} \left\{ \hat{L}_{(X,y)}(h) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) \right\}$

Empirical risk

Generalization gap & **sample complexity**:

$$L\left(\hat{h}_{(X,y)}\right) - L(h^*) \leq \widetilde{O}\left(\sqrt{\frac{\text{Complexity}(\mathscr{L}, \mathscr{H})}{n}}\right)$$

with high probability over $(X, y) \sim P(x, y)^n$

Data Augmentation

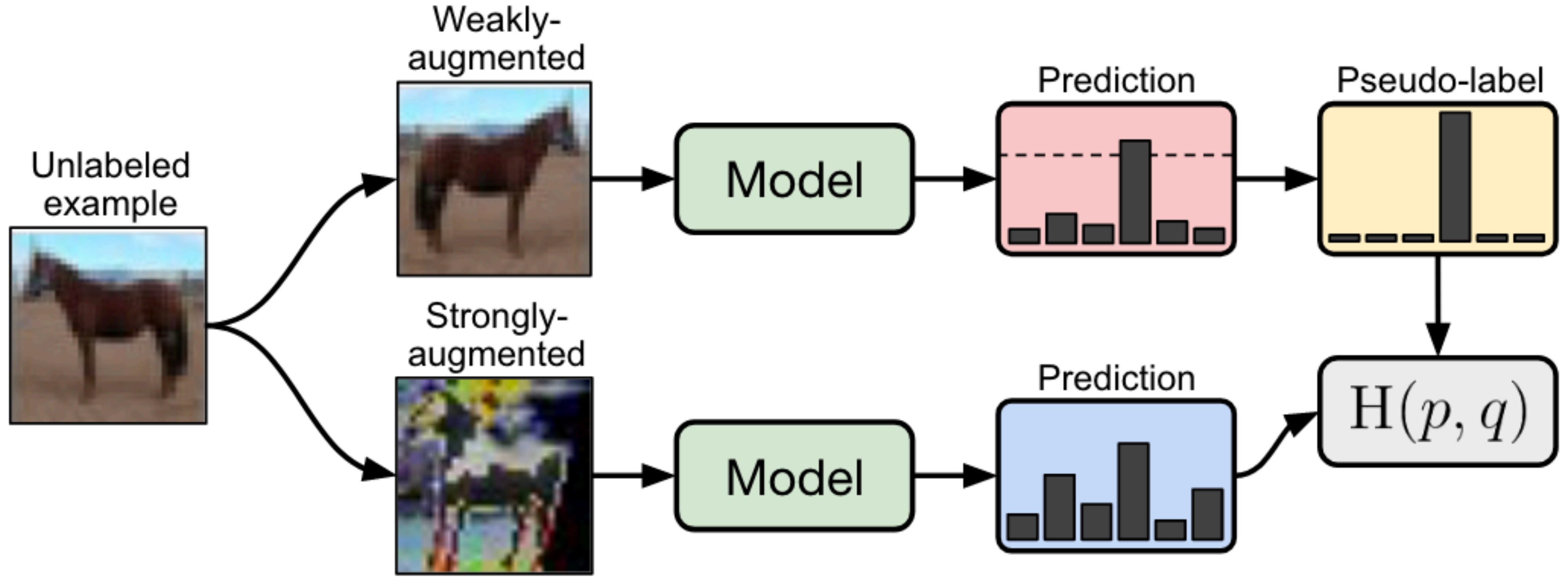
- A transformation $A : \mathscr{X} \to \mathscr{X}$ that **preserves semantic information** in $x \in \mathscr{X}$, e.g., random rotation, cropping, color jittering on images
- Consider an augmented training set of $(X, y) \sim P(x, y)^n$ with $\alpha \in \mathbb{N}$ augmentations $\left\{x_{i,j} = A_{i,j}(x_i)\right\}_{j\in[\alpha]}$ per sample $i \in [n]$: with $\mathbf{M} = \left[\mathbf{I}_n; \cdots; \mathbf{I}_n\right] \in \mathbb{R}^{(1+\alpha)n\times n}$ being the vertical stack of $n \times n$ identity matrices,

$$\left(\mathscr{A}(X), \mathbf{M}y\right) = \left(\left[x_1, \cdots, x_n, x_{1,1}, \cdots, x_{n,1}, \cdots, x_{1,\alpha}, \cdots, x_{n,\alpha}\right]^\top, \left[y; y; \cdots; y\right]\right) \in \mathscr{X}^{(1+\alpha)n} \times \mathscr{Y}^{(1+\alpha)n},$$

- Proper data augmentations lead to **better generalization and sample complexity**
- Ubiquitous in SOTA methods, with diverse designs (e.g., Mixup, Cutout, RandAugment, UDA, etc.)

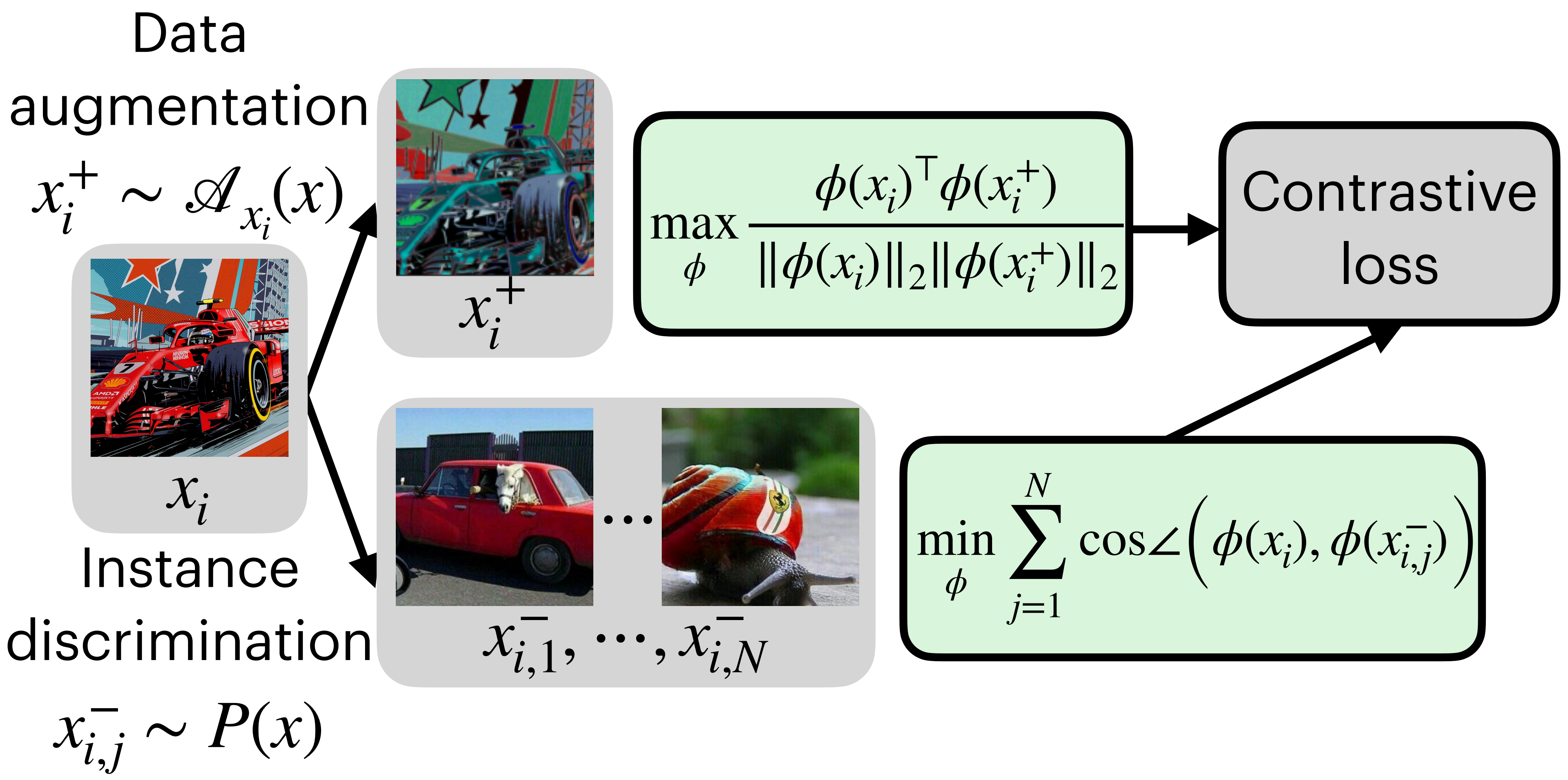
Data Augmentation

- Semi-supervised learning
  - **Data augmentation consistency (DAC) regularization**
  - E.g., MixMatch (Berthelot et al, 2019), Fixmatch (Sohn et al, 2020)



Semi-supervised learning with DAC regularization via FixMatch (Sohn et al, 2020, Fig. 1)

- Self-supervised learning
  - Contrastive learning
  - E.g., MoCo (He et al, 2020), SimCLR (Chen et al, 2020)



Sources of potency?

- Large amounts of unlabeled data
- **Effective algorithms for utilizing data augmentations**

Algorithmic Choices of Leveraging Data Augmentation in **Supervised Learning**

DA-ERM: ERM on augmented training samples $\left(\mathscr{A}(X), \mathbf{M}y\right)$

- $$\widehat{h}^{\text{DA-ERM}}_{(X,y)} \triangleq \underset{h \in \mathscr{H}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(h(x_i), y_i) + \sum_{i=1}^{n} \sum_{j=1}^{\alpha} \ell(h(x_{i,j}), y_i)$$

DAC: data augmentation consistency regularization

- $$\widehat{h}^{\text{DAC}}_{(X,y)} \triangleq \underset{h \in \mathscr{H}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(h(x_i), y_i) + \lambda \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{\alpha} \varrho(\phi_h(x_{i,j}), \phi_h(x_i))}_{\text{DAC regularization}}$$

- $\phi_h : \mathscr{X} \to \mathscr{W}$ is a representation function associated with $h \in \mathscr{H}$
  - $\mathscr{W}$ is a (latent) metric space with metric $\varrho : \mathscr{W} \times \mathscr{W} \to \mathbb{R}_{\geq 0}$
  - $h = f_h \circ \phi_h$ where $\phi_h(x)$ encapsulates semantic information in x
  - E.g., ϕ_h = neural network, f_h = linear classifier, $\varrho(u, v) = \|u - v\|_2$

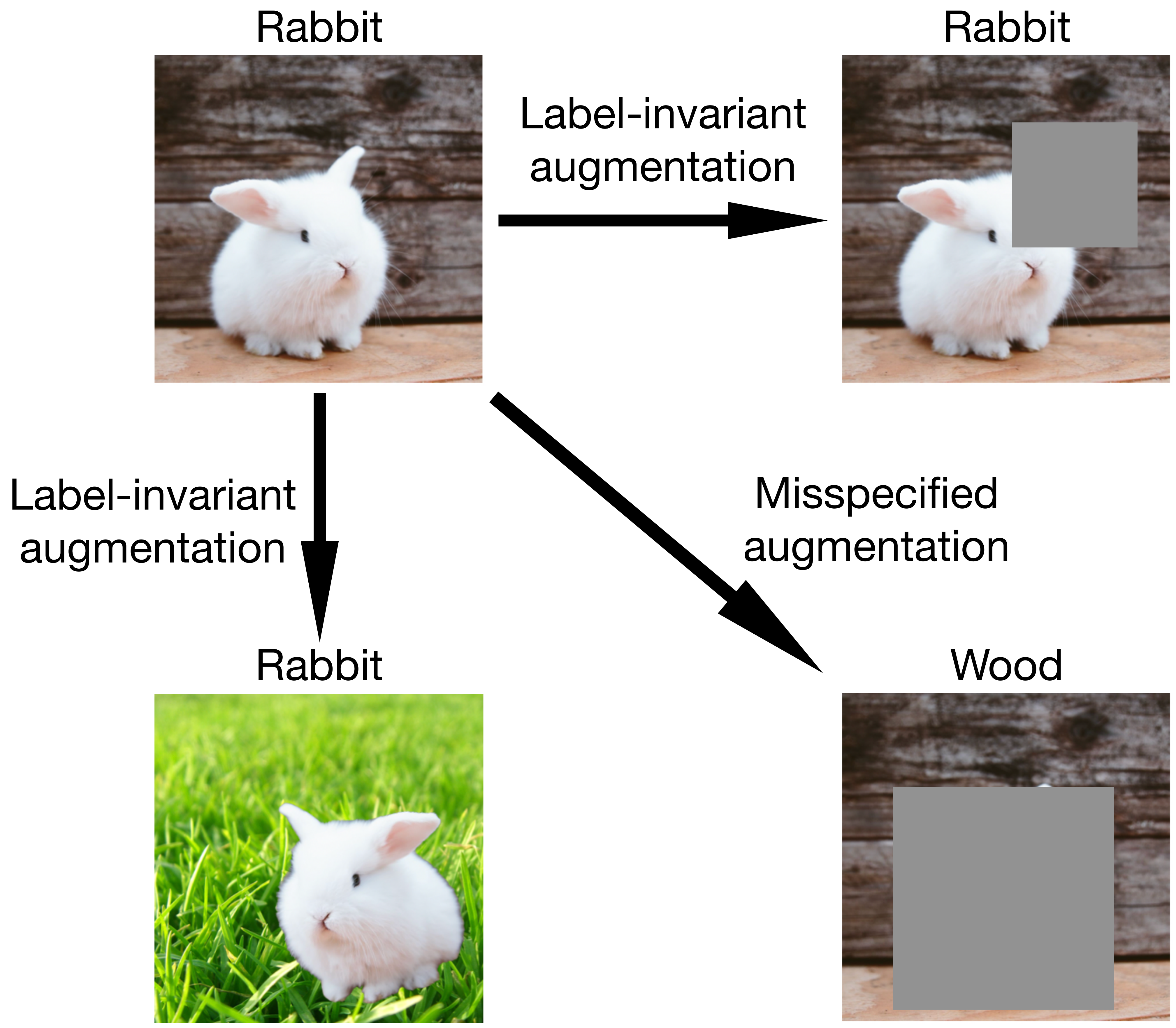
Whether potency of DAC comes merely from unlabeled data, or

DAC has intrinsic algorithmic advantage over DA-ERM?

- Apple-to-apple comparisons in **supervised learning setting**
- With limited random augmentations
- With “good”/“bad” augmentations
- From linear model to neural network

Label-invariant ("Good") v.s. Misspecified ("Bad") Data Augmentations

- Label-invariant ("good") augmentation
  - Augmentation preserves labels: $P(y \mid x) = P(y \mid A(x))$
  - $\varrho(\phi_{h*}(x_{i,j}), \phi_{h*}(x_i)) = 0$ for all $i \in [n], j \in [\alpha]$
- Misspecified ("bad") augmentation
  - Augmentation perturbs labels: $P(y \mid x) \neq P(y \mid A(x))$
  - $0 < \sum_{i=1}^{n} \sum_{j=1}^{\alpha} \varrho(\phi_{h*}(x_{i,j}), \phi_{h*}(x_i)) < C_{mis}$



Linear Regression + Label-invariant Augmentation

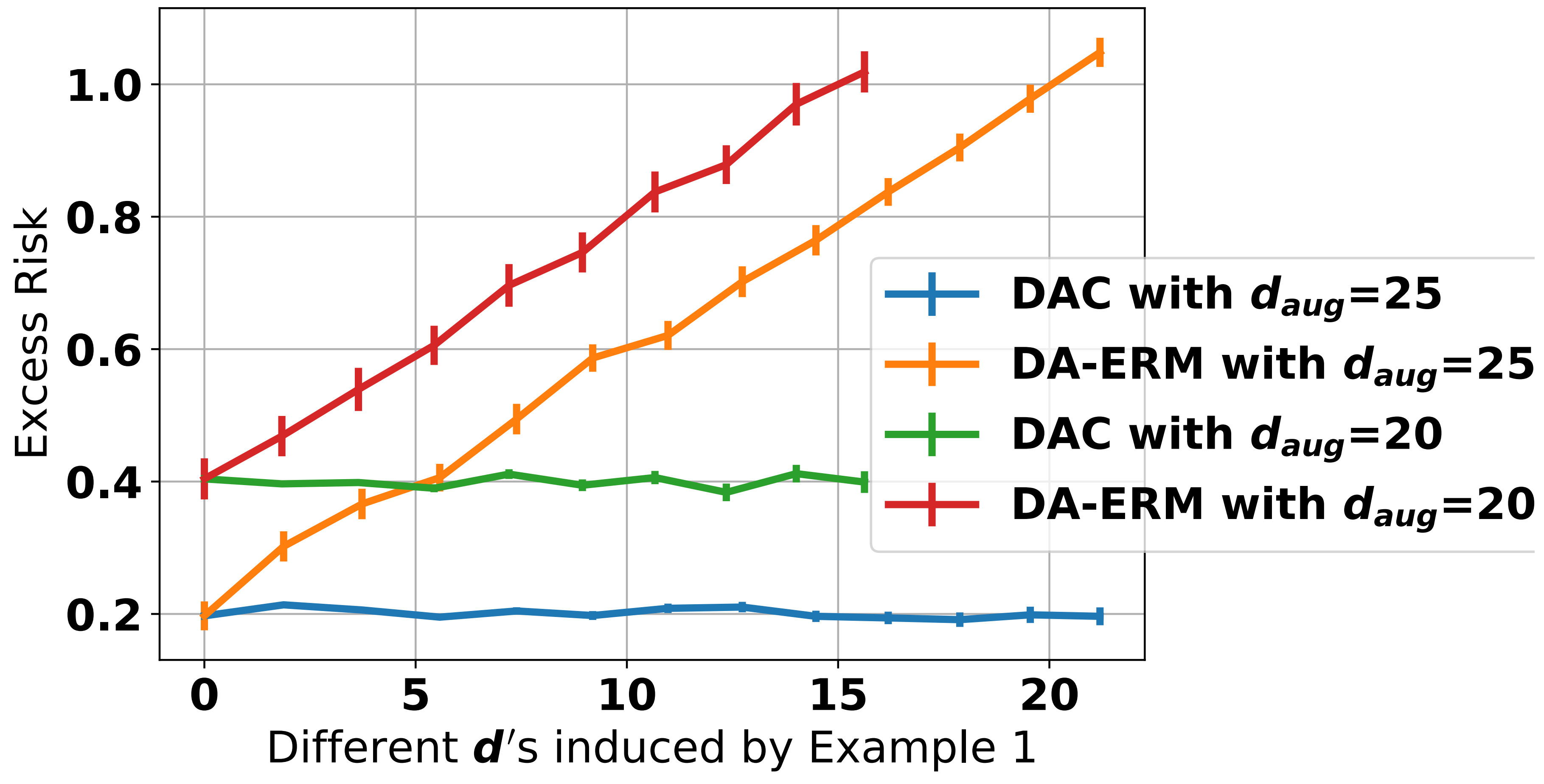
- Dimension-d linear regression: $(\underset{n\times d}{X}, y) = \left\{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\right\}_{i=1}^{n}$ with $y = X\theta^* + \epsilon$, $\epsilon \sim \mathcal{N}(0,\sigma^2\mathbf{I}_n)$
- $\theta^* = \underset{\theta\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ L(\theta) \triangleq \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I}_n)}\left[\frac{1}{n}\|y - X\theta\|_2^2\right] \right\}$, $\widehat{\theta}^{ERM} = \underset{\theta\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \widehat{L}(\theta) \triangleq \frac{1}{n}\|y - X\theta\|_2^2 \right\}$
- Without data augmentation: assume $\operatorname{rank}(X) = d$, $\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I}_n)}\left[L\left(\widehat{\theta}^{ERM}\right) - L\left(\theta^*\right)\right] = \frac{d\sigma^2}{n}$
- **Label-invariant augmentations**: $\Delta \triangleq \mathscr{A}(X) - \mathbf{M}X \in \mathbb{R}^{n\times d}$ such that $\Delta\theta^* = 0$
- **Augmentation strength**: $d_{aug} \triangleq \operatorname{rank}(\Delta)$ such that larger $d_{aug} \Rightarrow$ stronger augmentation
- Assume $\operatorname{rank}(\mathscr{A}(X)) = d$, taking $\lambda \to \infty$ (i.e., **DAC constraint** $\Delta\theta = 0$), with $d' \triangleq \frac{1}{1+\alpha}\operatorname{tr}\left(\left(\mathscr{A}(X)\mathscr{A}(X)^\dagger - \mathbf{P}_{\mathcal{S}}\right)\mathbf{M}\mathbf{M}^\top\right) \in [0, d_{aug}]$ where $\mathbf{P}_{\mathcal{S}}$ denotes the orthogonal projector onto $\mathcal{S} \triangleq \left\{\mathbf{M}X\theta \mid \Delta\theta = 0\right\}$:

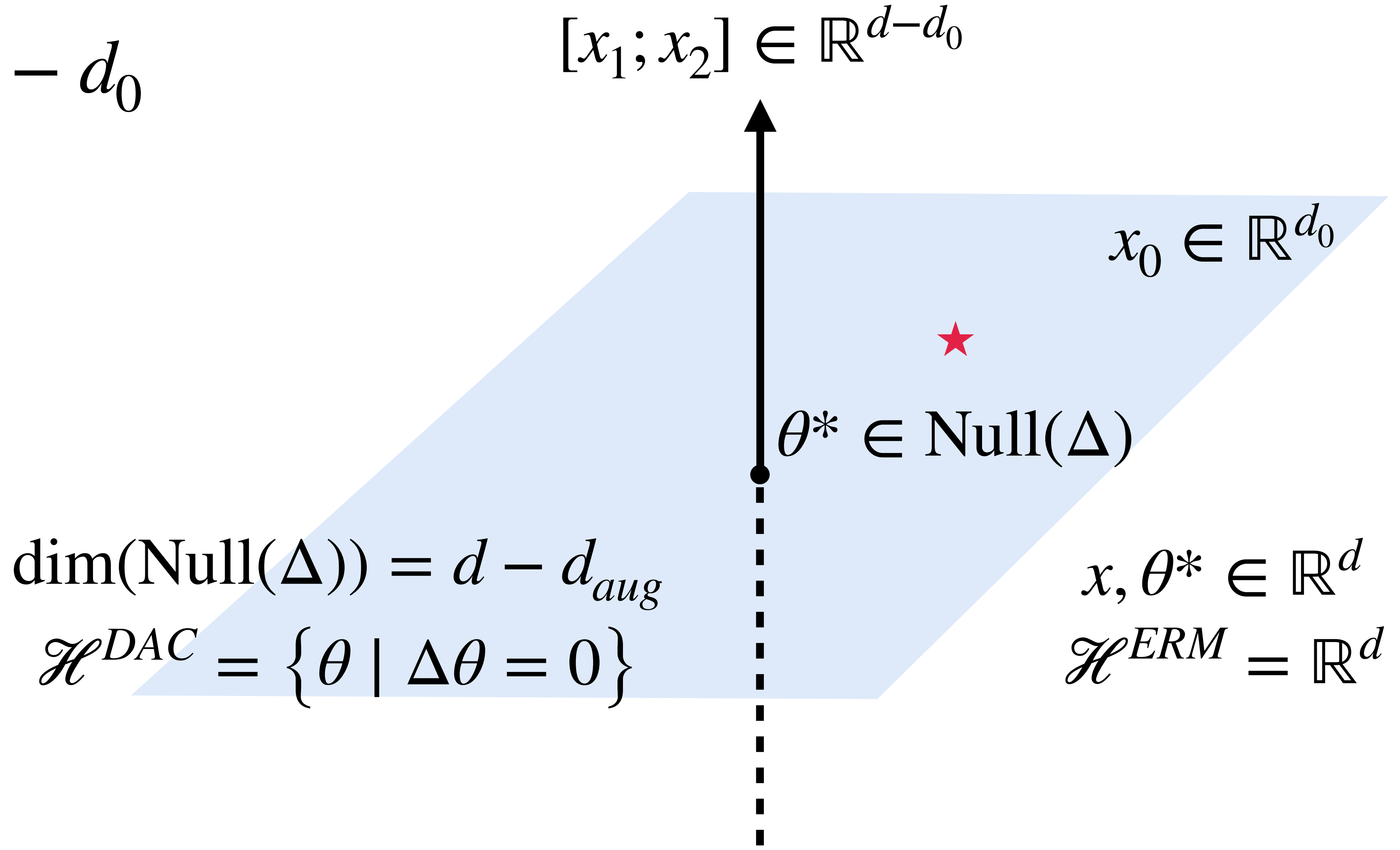
$$\mathbb{E}_\epsilon\left[L\left(\widehat{\theta}^{DAC}\right) - L\left(\theta^*\right)\right] = \frac{(d - d_{aug})\sigma^2}{n} \qquad \mathbb{E}_\epsilon\left[L\left(\widehat{\theta}^{DA-ERM}\right) - L\left(\theta^*\right)\right] = \frac{(d - d_{aug} + d')\sigma^2}{n}$$

Linear Regression + Label-invariant Augmentation

Example 1

- Ground truth: $\theta^* = [\theta_1, \cdots, \theta_{d_0}, 0, \cdots, 0]$ with $\theta_i \sim \mathcal{N}(0,1)\ \forall\ i \in [d_0]$
- Semantic subspace $d_0 < d$ & spurious subspace $d - d_0 = d_1 + d_2$
- Data: $n = 50$, $d = 30$, $P(x) = \mathcal{N}(0, \mathbf{I}_d)$ such that $x = [\underset{d_0}{x_0}; \underset{d_1}{x_1}; \underset{d_2}{x_2}]$, $\sigma = 1$
- Augmentation: $A([x_0; x_1; x_2]) = [x_0; 2x_1; -x_2]$ such that $d_{aug} = d - d_0$





Beyond Label-invariant Augmentation

- **Misspecified augmentations**: $\Delta = \mathscr{A}(X) - \mathbf{M}X \in \mathbb{R}^{n \times d}$ whereas $\Delta\theta^* \neq 0$
- Recall $d_{aug} = \text{rank}(\Delta)$, denote $\mathbf{P}_\Delta = \Delta^\dagger \Delta$, $\widetilde{\Delta} = \left(\mathbf{M}X\mathscr{A}(X)^\dagger\right)\Delta$, $S = \frac{1}{1+\alpha}\mathbf{M}^\top \mathscr{A}(X)$
- Let $\Sigma_X = \frac{1}{n}X^\top X$, $\Sigma_{\mathscr{A}(X)} = \frac{1}{(1+\alpha)n}\mathscr{A}(X)^\top \mathscr{A}(X)$, $\Sigma_\Delta = \frac{1}{(1+\alpha)n}\Delta^\top\Delta$, $\Sigma_{\widetilde{\Delta}} = \frac{1}{(1+\alpha)n}\widetilde{\Delta}^\top\widetilde{\Delta}$, $\Sigma_S = \frac{1}{n}S^\top S$
- Assume there exist $c_X, c_S > 0$ such that $\Sigma_{\mathscr{A}(X)} \preccurlyeq c_X \Sigma_X$ and $\Sigma_{\mathscr{A}(X)} \preccurlyeq c_S \Sigma_S$. Then, with **the optimal choice of** λ:

$$\mathbb{E}_\epsilon\left[L\left(\widehat{\theta}^{DAC}\right) - L(\theta^*)\right] \leq \frac{(d-d_{aug})\sigma^2}{n} + \left\| \mathbf{P}_\Delta\theta^* \right\|_{\Sigma_\Delta} \sqrt{\frac{\sigma^2}{n}\text{tr}\left(\Sigma_X \Sigma_\Delta^\dagger\right)}$$

$$\mathbb{E}_\epsilon\left[L\left(\widehat{\theta}^{DA-ERM}\right) - L(\theta^*)\right] \geq \frac{d\sigma^2}{n \cdot c_X c_S} + \left\| \mathbf{P}_\Delta\theta^* \right\|^2_{\Sigma_{\widetilde{\Delta}}}$$

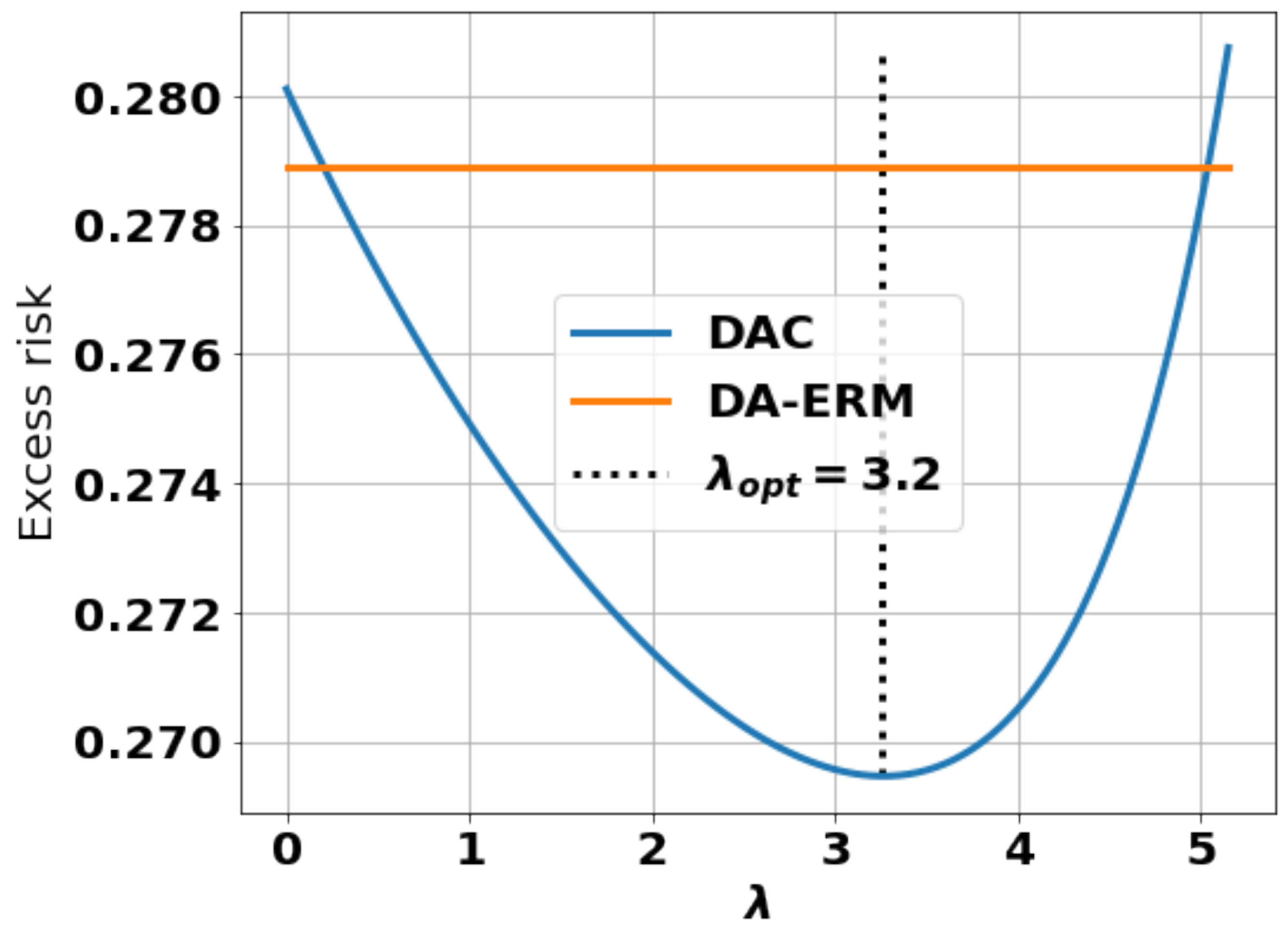
Variance   Bias

- $\mathbf{P}_\Delta\theta^*$ measures misspecification of augmentations $\mathscr{A}(X)$ in θ^*
- For DA-ERM, the bias term $\left\| \mathbf{P}_\Delta\theta^* \right\|^2_{\Sigma_{\widetilde{\Delta}}}$ induced by misspecification $\Delta\theta^* \neq 0$ fails to vanish as $n \to \infty$

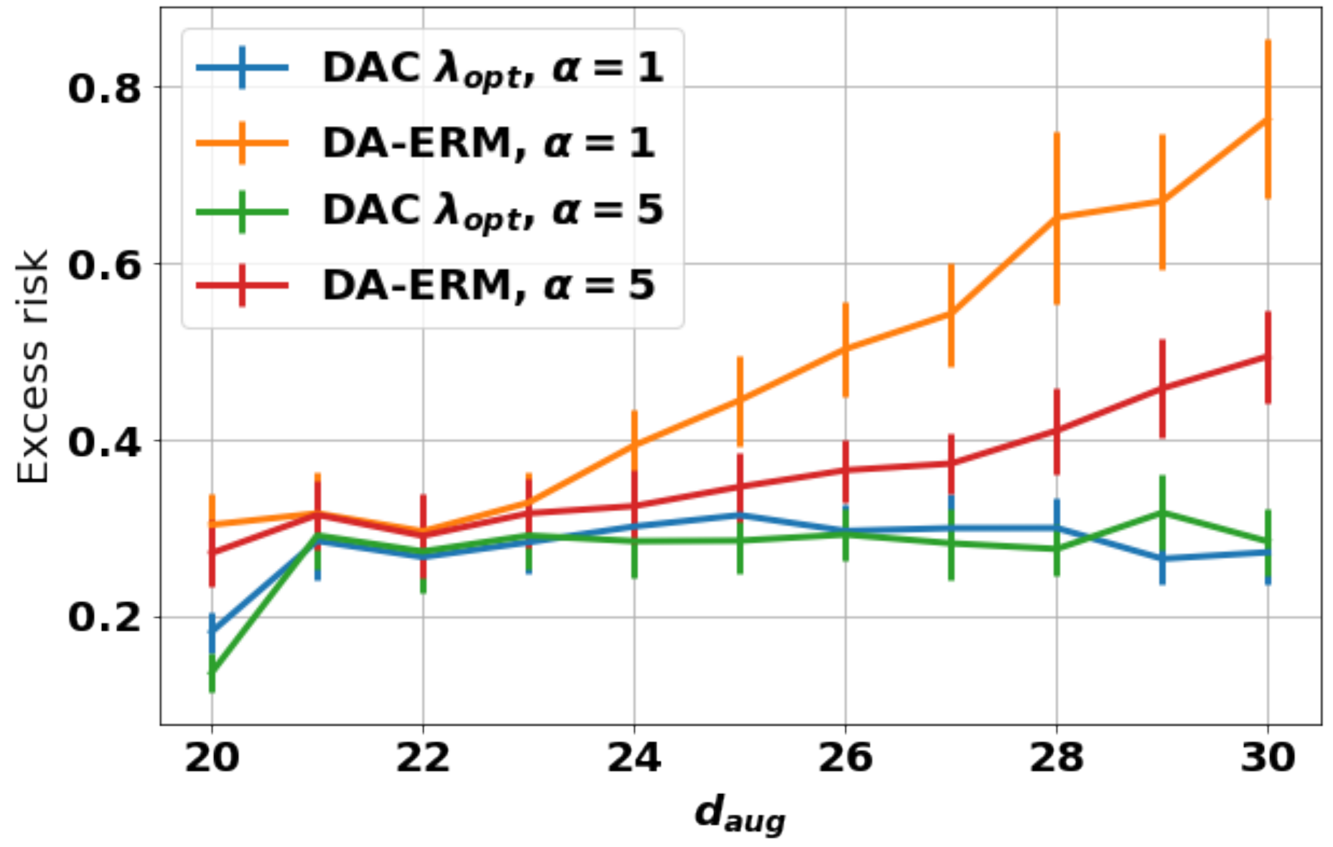
Beyond Label-invariant Augmentation

Example 2

- Ground truth: $\theta^* = [\theta_1, \cdots, \theta_{d_0}, 0, \cdots, 0]$ with $\theta_i \sim \{\pm 1\}\ \forall\ i \in [d_0]$, $d_0 < d$
- Data: $n = 50$, $d = 30$, $d_0 = 10$, $P(x) = \mathcal{N}(0, \mathbf{I}_d)$ such that $x = [\underbrace{x_0}_{d-d_{aug}} ; \underbrace{x_1}_{d_{aug}}]$, $\sigma = 0.1$
- Augmentation: $A([x_0; x_1]) = [x_0; x_1 + x_1']$ with $x_1' \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{d_{aug}}) \Rightarrow \mathbb{P}\left[\text{rank}(\Delta) = d_{aug}\right] = 1$
- Misspecification in ground truth: $\left[\theta_{d-d_{aug}+1}, \cdots, \theta_{d_0}\right]$



- The optimal λ implicitly incorporates knowledge on the upper bound of misspecification $\left\| \mathbf{P}_\Delta \theta^* \right\|_{\Sigma_\Delta}$
- Less misspecification $\Rightarrow$ larger λ



- Misspecified dimension: $d_{aug} - d_0$
- DAC is more robust to misspecification (with larger d_{aug})
- DAC leverages augmentations more efficiently (with smaller α)

Beyond Linear Model

- Two-layer ReLU network regression
  - $P(x, y)$: $y = h^*(x) + \epsilon$ where $\epsilon \sim \mathcal{N}(0,\sigma^2)$ and $h^*(x) = \max\left(x^\top B^*, 0\right) w^*$, $w^* \in \mathbb{R}^q$, $B^* = \left[b_1^*, \cdots, b_q^*\right] \in \mathbb{R}^{d\times q}$
  - Bounded ground truth: $\|w^*\|_1 \le C_w$, $\|b_j^*\|_2 = 1 \ \forall\, j \in [q]$
  - Function class: $\mathcal{H} = \left\{\max(\cdot^\top B, 0) w \mid B = [b_1, \cdots, b_q],\ \|b_j\|_2 = 1 \ \forall\, j \in [q],\ \|w\|_1 \le C_w\right\} \ni h^*$
- DAC constraint on hidden layer: $\max\left(\mathcal{A}(X)B, 0\right) = \max\left(\mathbf{M}XB, 0\right)$
- Recall $\Delta = \mathcal{A}(X) - \mathbf{M}X \in \mathbb{R}^{n\times d}$ such that $\Delta\theta^* = 0$ and $d_{aug} \triangleq \mathrm{rank}(\Delta)$; $\mathbf{P}_\Delta^\perp = \mathbf{I}_d - \Delta^\dagger\Delta$
- Under mild regularity conditions: $\alpha n \ge 3d_{aug}$; $P(x)$ is zero-mean and subgaussian; Δ admits absolutely continuous distribution
- Conditioned on $X \sim P(x)^n$ and random augmentations Δ such that $\sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{P}_\Delta^\perp x_i\|_2^2} \le C_N$, with probability $\ge 1 - \delta$ over $P(y \mid x)$

$$L\left(\hat{\theta}^{DAC}\right) - L(\theta^*) \lesssim \sigma C_w \left(\frac{C_N}{\sqrt{n}} + C_N\sqrt{\frac{\log(1-\delta)}{n}}\right)$$

$$L\left(\hat{\theta}^{DA-ERM}\right) - L(\theta^*) \lesssim \sigma C_w \max\left(\sqrt{\frac{d-d_{aug}}{n}}, \sqrt{\frac{d}{(1+\alpha)n}}\right)$$

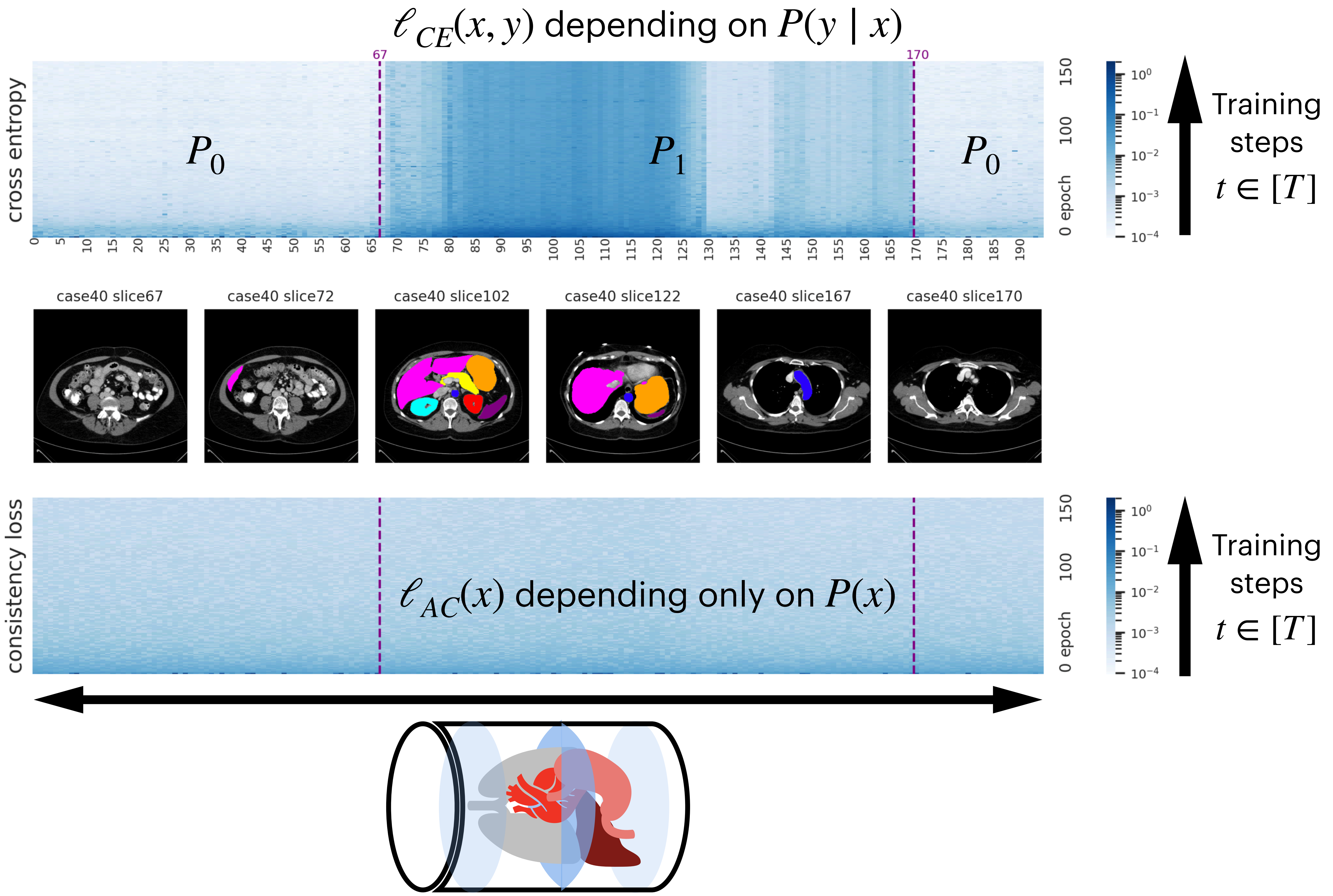
- Randomness in $X \sim P(x)^n$: with a sufficiently large n, $C_N \le \sqrt{d - d_{aug}}$ with high probability

Adaptively Weighted Data Augmentation Consistency Regularization for Distributionally Robust Optimization under Concept Shift

Based on joint work with: Yuege Xie, Rachel Ward

Dong Y, Xie Y, Ward R. AdaWAC: Adaptively Weighted Augmentation Consistency Regularization for Volumetric Medical Image Segmentation. arXiv preprint arXiv:2210.01891. 2022 Oct 4.

Information Imbalance in Medical Image Segmentation



- Input image: $x \in \mathcal{X} \subseteq \mathbb{R}^d$
- Segmentation label: $y \in [K]^d$
- Ground truth distribution $P_\xi : \mathcal{X} \times \mathcal{Y} \to [0,1]$
- **Information imbalance**: $P_\xi = \xi P_0 + (1 - \xi) P_1$
  - P_0, P_1 with disjoint supports $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$
  - P_0: label-sparse distribution
  - P_1: label-dense distribution
  - $P_0(x) = P_1(x)$ uniform for all $x \in \mathcal{X}$
  - **Concept shift**: $P_0(y \mid x) \neq P_1(y \mid x)$

How to improve segmentation accuracy under such concept shift?

Concept Shift: Label-sparse v.s. Label-dense Samples

- Data augmentation: $A_{i,j} \sim \mathscr{A}^{2n}\ \forall\ i \in [n],\ j \in [2]$; $\mathscr{A}$ is a distribution over mild random augmentations (i.e., rotation & flip)
- Class of segmentation functions: $\mathscr{F} = \left\{ f_\theta = \underbrace{\psi_\theta}_{\text{decoder}} \circ \underbrace{\phi_\theta}_{\text{encoder}} \mid \theta \in \mathscr{F}_\theta,\ \phi_\theta : \mathscr{X} \to \mathscr{Z} \subseteq \mathbb{R}^q,\ \psi_\theta : \mathscr{Z} \to \mathbb{R}^{d\times K} \right\}$
  - Ground truth $\theta^* = \cap_{\xi\in[0,1]} \operatorname*{argmin}_{\theta\in\mathscr{F}_\theta} \mathbb{E}_{P_\xi}\left[\ell_{CE}(\theta;(x,y))\right]$, Banach space $\left(\mathscr{F}_\theta, \|\cdot\|_{\mathscr{F}}\right)$
- Supervised cross-entropy loss: $\ell_{CE}\left(\theta;(x,y)\right) = -\frac{1}{d}\sum_{j=1}^{d}\sum_{k=1}^{K} \mathbb{I}\{y_j = k\} \cdot \log\left(f_\theta(x)_{j,k}\right)$
- Unsupervised consistency regularization: $\ell_{AC}\left(\theta; x, A_1, A_2\right) = \lambda_{AC} \cdot \left\| \phi_\theta\left(A_1(x)\right) - \phi_\theta\left(A_2(x)\right) \right\|_2$

Assumption. (n-separation between label-sparse and label-dense distributions)

Given $\gamma > 0$, let $\mathscr{F}_{\theta^*}(\gamma) = \left\{\theta \in \mathscr{F}_\theta \mid \|\theta - \theta^*\|_{\mathscr{F}} \leq \gamma\right\}$ be a compact and convex neighborhood with pre-trained segmentation functions. We say that the label-sparse distribution P_0 and label-dense distribution P_1 are **n-separated over $\mathscr{F}_{\theta^*}(\gamma)$** if there exist $\omega > 0$ such that, with probability $\geq 1 - \Omega\left(n^{1+\omega}\right)$ over $\left(x, y, A_1, A_2\right) \sim \mathscr{X}\times\mathscr{Y}\times\mathscr{A}\times\mathscr{A}$: for all $\theta \in \mathscr{F}_{\theta^*}(\gamma)$,

$$(x,y) \sim P_0 \Rightarrow \ell_{CE}\left(\theta;(x,y)\right) < \ell_{AC}\left(\theta;x,A_1,A_2\right) \quad \& \quad (x,y) \sim P_1 \Rightarrow \ell_{CE}\left(\theta;(x,y)\right) > \ell_{AC}\left(\theta;x,A_1,A_2\right)$$

Sample Reweighting & Data Augmentation Consistency Regularization

- Both sample reweighting and consistency regularization are known for boosting distributional robustness
- Sample reweighting via **distributionally robust optimization (DRO)**

$$\min_{\theta} \max_{i \in [n]} \ell_{CE}\left(\theta; (x_i, y_i)\right) \Leftrightarrow \min_{\theta} \max_{\beta \in \Delta_n} \frac{1}{n} \sum_{i=1}^{n} \beta_i \cdot \ell_{CE}\left(\theta; (x_i, y_i)\right)$$

- Data augmentation **consistency regularization**

$$\min_{\theta} \ell_{CE}\left(\theta; (x_i, y_i)\right) + \ell_{AC}\left(\theta; x, A_1, A_2\right)$$

- A key challenge of incorporating consistency regularization in medical image segmentation
  - Dense segmentation labels are sensitive to data augmentations (even the simplest ones like rotation and flip)
  - For label-dense samples, severe misspecification is inevitable for data augmentations
  - For label-sparse samples, misspecification is mild (if any) thanks to the sparsity of labeled pixels
- How to properly combine DRO and consistency regularization for better distributional robustness?

Weighted Data Augmentation Consistency (WAC) Regularization

$$\widehat{\theta}, \widehat{\beta} = \underset{\theta \in \mathcal{F}_{\theta^*}(\gamma)}{\operatorname{argmin}} \ \underset{\beta \in [0,1]^n}{\operatorname{argmax}} \left\{ \widehat{L}^{WAC}(\theta, \beta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \beta_i \cdot \ell_{CE}\left(\theta; (x_i, y_i)\right) + (1 - \beta_i) \cdot \ell_{AC}\left(\theta; x_i, A_{i,1}, A_{i,2}\right) \right\}$$

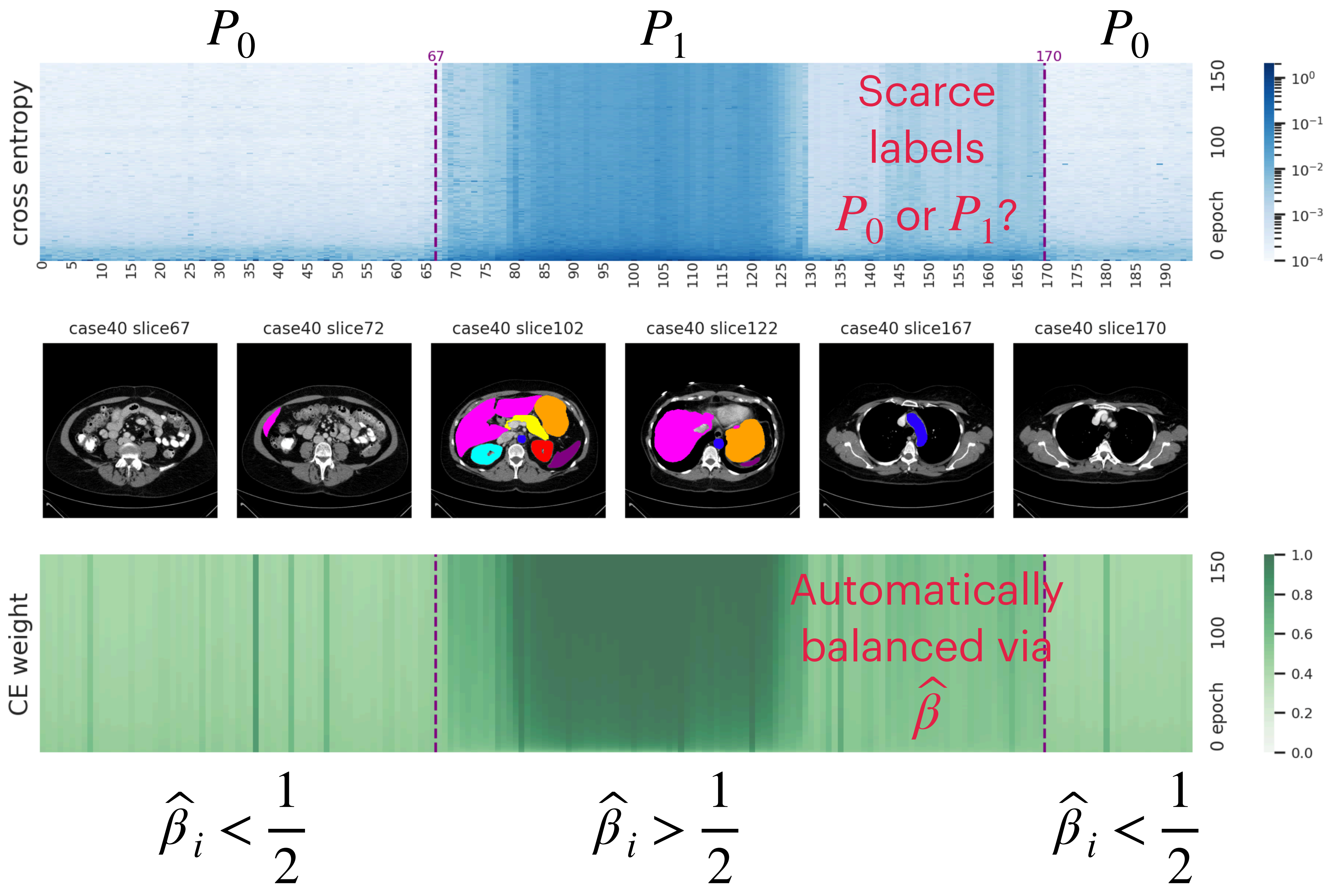
Proposition. (Separation of P_0 and P_1 at saddle point)

Assume that $\ell_{CE}\left(\theta; (x, y)\right)$ and $\ell_{AC}\left(\theta; x, A_1, A_2\right)$ are convex and continuous in θ for all $\left(x, y, A_1, A_2\right)$; recall $\mathcal{F}_{\theta^*}(\gamma)$ is convex and compact. If P_0 and P_1 are n-separated, there exist $\widehat{\beta} \in \{0,1\}^n$ and $\widehat{\theta} \in \underset{\theta \in \mathcal{F}_{\theta^*}(\gamma)}{\operatorname{argmin}} \widehat{L}^{WAC}(\theta, \widehat{\beta})$ such that

$$\underset{\theta \in \mathcal{F}_{\theta^*}(\gamma)}{\operatorname{argmin}} \widehat{L}^{WAC}(\theta, \widehat{\beta}) = \widehat{L}^{WAC}(\widehat{\theta}, \widehat{\beta}) = \underset{\beta \in [0,1]^n}{\operatorname{argmax}} \widehat{L}^{WAC}(\widehat{\theta}, \beta)$$

where $\widehat{\beta}$ separates label-sparse and label-dense samples:

$$\widehat{\beta}_i = \begin{cases} 0, & (x_i, y_i) \sim P_0 \\ 1, & (x_i, y_i) \sim P_1 \end{cases}$$



AdaWAC: Adaptively Weighted Augmentation Consistency Regularization

Online mirror descent for saddle point problem

- $\max_{B \in \Delta_2^n}$ via mirror map $\varphi_B(B) = \sum_{i=1}^{n} \sum_{j=1}^{2} B_{ij} \log\left(B_{ij}\right)$
- $\min_{\theta}$ via gradient descent $\varphi_\theta(\theta) = \| \theta - \theta^* \|_{\mathcal{F}}^2$

1. Initialize weights $\beta^{(0)} = (1/2, \cdots, 1/2) \in [0,1]^n$
2. For $t = 0, \cdots, T$:
   1. Sample $i_t \sim [n]$ uniformly; set $b \leftarrow \left[\beta_{i_t}^{(t-1)}, 1 - \beta_{i_t}^{(t-1)}\right]$, $\beta^{(t)} \leftarrow \beta^{(t-1)}$
   2. Update $b_1 \leftarrow b_1 \cdot \exp\left(\eta_\beta \cdot \ell_{CE}\left(\theta^{(t-1)}; (x_{i_t}, y_{i_t})\right)\right)$, $b_2 \leftarrow b_2 \cdot \exp\left(\eta_\beta \cdot \ell_{AC}\left(\theta^{(t-1)}; x_{i_t}, A_{i_t,1}, A_{i_t,2}\right)\right)$, $\beta_{i_t}^{(t)} \leftarrow \frac{b_1}{\|b\|_1}$
   3. Update $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta_\theta \cdot \left(\beta_{i_t}^{(t)} \cdot \nabla_\theta \ell_{CE}\left(\theta^{(t-1)}; (x_{i_t}, y_{i_t})\right) + \left(1 - \beta_{i_t}^{(t)}\right) \cdot \nabla_\theta \ell_{AC}\left(\theta^{(t-1)}; x_{i_t}, A_{i_t,1}, A_{i_t,2}\right)\right)$

Proposition. (Convergence of AdaWAC)

If there exist $C_\theta, C_\beta > 0$ such that $\forall\, \theta \in \mathcal{F}_{\theta^*}(\gamma)$, $\beta \in [0,1]^n$, $\frac{1}{n} \sum_{i=1}^{n} \max\left\{\ell_{CE}\left(\theta; (x_i, y_i)\right), \ell_{AC}\left(\theta; x_i, A_{i,1}, A_{i,2}\right)\right\}^2 \leq C_\beta^2$ and $\frac{1}{n} \sum_{i=1}^{n} \left\| \beta_i \cdot \nabla_\theta \ell_{CE}\left(\theta; (x_i, y_i)\right) + \left(1 - \beta_i\right) \cdot \nabla_\theta \ell_{AC}\left(\theta; x_i, A_{i,1}, A_{i,2}\right) \right\|_{\mathcal{F}}^2 \leq C_\theta^2$, then with $\eta_\theta = \eta_\beta = \frac{2}{\sqrt{5T\left(\gamma^2 C_\theta^2 + 2nC_\beta^2\right)}}$

$$\mathbb{E}\left[\max_{\beta \in [0,1]^n} \widehat{L}^{WAC}\left(\bar{\theta}_T, \beta\right) - \min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}\left(\theta, \bar{\beta}_T\right)\right] \leq 2\sqrt{5\left(\gamma^2 C_\theta^2 + 2nC_\beta^2\right)/T}$$

Sample Efficiency & Distributional Robustness of AdaWAC

AdaWAC v.s. baseline (ERM + SGD) with TransUNet on Synapse and its subsets

Training	Method	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
full	baseline	76.66 ± 0.88	29.23 ± 1.90	87.06	55.90	81.95	75.58	94.29	56.30	86.05	76.17
	AdaWAC	$\mathbf{79.04 \pm 0.21}$	$\mathbf{27.39 \pm 1.91}$	87.53	56.57	83.23	81.12	94.04	62.05	89.51	78.32
half-slice	baseline	74.62 ± 0.78	31.62 ± 8.37	86.14	44.23	79.09	78.46	93.50	55.78	84.54	75.24
	AdaWAC	$\mathbf{77.37 \pm 0.40}$	$\mathbf{29.56 \pm 1.09}$	86.89	55.96	82.15	78.63	94.34	57.36	86.60	77.05
half-vol	baseline	71.08 ± 0.90	46.83 ± 2.91	84.38	46.71	78.19	74.55	92.02	48.03	76.28	68.47
	AdaWAC	$\mathbf{73.81 \pm 0.94}$	$\mathbf{35.33 \pm 0.92}$	84.37	48.14	80.32	77.39	93.23	52.78	83.50	70.79
half-sparse	baseline	31.74 ± 2.78	69.72 ± 1.37	65.71	8.33	59.46	51.59	51.18	10.72	6.92	0.00
	AdaWAC	$\mathbf{41.03 \pm 2.12}$	$\mathbf{59.04 \pm 12.32}$	71.27	8.33	69.14	63.09	64.29	17.74	30.77	3.57

Sample efficiency

- **full**: original Synapse multi-organ dataset
- **half-slice**: slices with even indices in each case
- **half-vol**: 9 cases sampled uniformly from the total 18 training volumes

Distributional robustness

- **half-sparse**: the first half of slices in each volume, most of which are label-sparse

Comparison with Hard-thresholding algorithms

Why do we need adaptive weighting? Can we manually separate label-sparse & label-dense samples?

Comparison to hard-thresholding algorithms (+ consistency regularization) with TransUNet on Synapse

Method	baseline	trim-train	trim-train +ACR	trim-ratio	trim-ratio +ACR	pseudo-*AdaWAC*	*AdaWAC*
DSC ↑	76.66 ± 0.88	76.80 ± 1.13	78.42 ± 0.17	76.49 ± 0.16	77.71 ± 0.56	77.72 ± 0.65	$\mathbf{79.04 \pm 0.21}$
HD95 ↓	29.23 ± 1.90	32.05 ± 2.34	27.84 ± 1.16	31.96 ± 2.60	28.51 ± 2.66	28.45 ± 1.18	$\mathbf{27.39 \pm 1.91}$

- **trim-train** learns only from slices with at least one non-background pixel and trims the rest
- **trim-ratio** ranks the cross-entropy loss $\ell_{CE}(\theta; (x, y))$ in each iteration (mini-batch) and trims samples with the lowest $\ell_{CE}(\theta; (x, y))$ (i.e., label-sparse) at a fixed ratio — the ratio of all-background slices in the full training set (≈ 0.42), updating only those samples with the higher $\ell_{CE}(\theta; (x, y))$ (i.e., label-dense)
- **pseudo-AdaWAC** simulates the sample weights $\widehat{\beta}$ at the saddle point — learns via $\ell_{CE}(\theta; (x, y))$ on slices with at least one non-background pixel while via $\ell_{AC}(\theta; x, A_1, A_2)$ otherwise
- **+ACR** further incorporates the augmentation consistency regularization directly via $+\ell_{AC}(\theta; x, A_1, A_2)$

Ablation Study

Ablation study of AdaWAC with TransUNet on Synapse

Method	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
baseline	76.66 ± 0.88	29.23 ± 1.90	87.06	55.90	81.95	75.58	94.29	56.30	86.05	76.17
reweight-only	76.27 ± 0.42	32.66 ± 3.48	87.30	52.56	81.21	75.77	94.13	58.96	84.69	75.52
reweight-EM	76.83 ± 0.62	31.95 ± 2.64	87.33	54.16	82.20	76.00	93.84	58.59	86.35	76.16
ACR-only	78.01 ± 0.62	27.78 ± 2.80	87.51	58.79	83.39	79.26	94.70	58.99	86.02	75.43
AdaWAC-0.01	77.75 ± 0.23	28.02 ± 3.50	87.33	56.68	83.35	78.53	94.45	57.02	87.72	76.94
AdaWAC-1.0	$\mathbf{79.04 \pm 0.21}$	$\mathbf{27.39 \pm 1.91}$	87.53	56.57	83.23	81.12	94.04	62.05	89.51	78.32

On the influence of **consistency regularization**

- **reweight-only**: standard DRO following Sagawa et al, 2020
- **reweight-EM**: DRO + entropy maximization proposed in Fidon et al, 2021
- Reweighting alone brings little improvement compared to the baseline

On the influence of **sample reweighting**

- **ACR-only**: $\min_{\theta} \ell_{CE}\left(\theta; (x_i, y_i)\right) + \ell_{AC}\left(\theta; x, A_1, A_2\right)$
- **AdaWAC-0.01**: $\eta_\beta = 0.01$ with slow separation
- **AdaWAC-1.0**: $\eta_\beta = 1.0$ with (properly) rapid separation
- Proper reweighting brings additional boost

- Sagawa S, Koh PW, Hashimoto TB, Liang P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731. 2019 Nov 20.
- Fidon L, Aertsen M, Mufti N, Deprest T, Emam D, Guffens F, Schwartz E, Ebner M, Prayer D, Kasprian G, David AL. Distributionally robust segmentation of abnormal fetal brain 3D MRI. InUncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3 2021 (pp. 263-273). Springer International Publishing.

Recent & Ongoing Works

Randomized numerical linear algebra

- Yijun Dong, Per-Gunnar Martinsson. "Simpler is better: A comparative study of randomized algorithms for computing the CUR decomposition". arXiv preprint arXiv:2104.05877. (2021). ACOM 2023
- Yijun Dong, Per-Gunnar Martinsson, Yuji Nakatsukasa. "Efficient Bounds and Estimates for Canonical Angles in Randomized Subspace Approximations". arXiv preprint arXiv:2211.04676. (2022).
- Ongoing work joint with Kate Pearce, Chao Chen, Per-Gunnar Martinsson: Randomized pivoting-based interpolative decomposition with efficient residual estimation

Statistical learning theory

- Shuo Yang*, Yijun Dong*, Rachel Ward, Inderjit S Dhillon, Sujay Sanghavi, Qi Lei. "Sample Efficiency of Data Augmentation Consistency Regularization". arXiv preprint arXiv:2202.12230. (2022). AISTATS 2023
- Yijun Dong*, Yuege Xie*, Rachel Ward. "AdaWAC: Adaptively Weighted Augmentation Consistency Regularization for Volumetric Medical Image Segmentation". arXiv preprint arXiv:2210.01891. (2022).
- Ongoing work joint with Kevin Miller, Qi Lei, Rachel Ward: Semi-supervised relational knowledge distillation as provable label propagation

Acknowledgements

Advisors: Prof. Per-Gunnar Martinsson, Prof. Rachel Ward

Thesis committee: Prof. George Biros, Prof. Joseph Kileel, Prof. Yuji Nakatsukasa

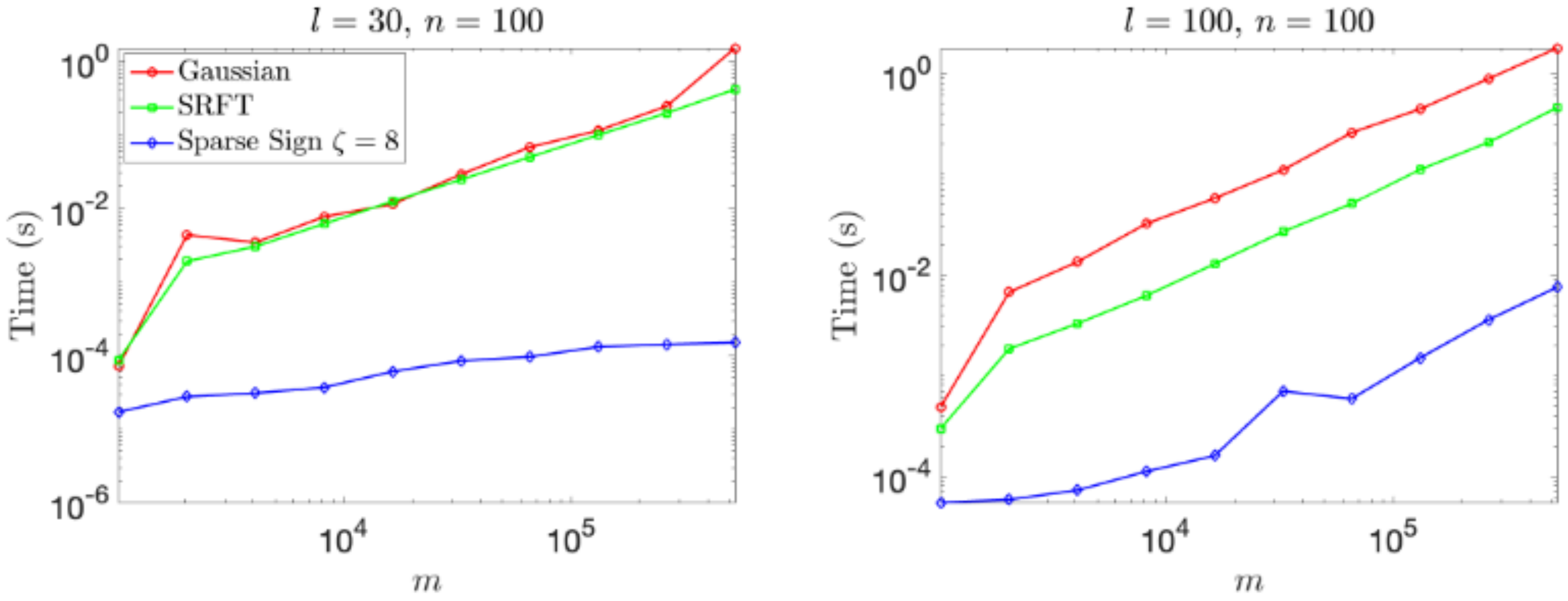
Collaborators: Shuo Yang, Prof. Inderjit Dhillon, Prof. Sujay Sanghavi, Prof. Qi Lei, Dr. Yuege Xie, Dr. Kevin Miller, Dr. Kate Pearce, Dr. Chao Chen

And many more

Thank You!

Appendix

Efficiency of Sketching



- Runtime: Sparse sign $O(mn\zeta)$ < SRFT $O(mn \log l)$ < **Gaussian** $O(mnl)$
- Efficiency of sketching is important (only) when l is sufficiently large
- Similar low-rank approximation errors $\|\mathbf{A} - \mathbf{A}\mathbf{X}^{\dagger}\mathbf{X}\|$ in practice
- We focus on **Gaussian embedding** for simplicity & consistency

Runtime of Gaussian / SRFT / sparse sign embedding $\boldsymbol{\Gamma} \in \mathbb{C}^{l \times m}$ on dense matrices of size $m \times n$