Beyond Real Data: Synthetic Data through the lens of Regularization

Amitis Shidani, Tyler Farghly, Yang Sun, Habib Ganjgahi* & George Deligiannidis*
* co-supervision

AISTATS 2026

Real data is scarce, but synthetic data is a double-edged sword

- In many critical domains like healthcare, collecting real data is challenging:
  - Expensive
  - Slow
  - Constrained by privacy
- Generative models like diffusion models can produce synthetic samples

There is a fundamental tension:

Too little synthetic data = missed opportunity
Too much = distributional mismatch degrades performance

The central question

What is the optimal ratio of synthetic to real data that minimizes generalization error?

- Learning-theoretic framework (kernel regression + stability)
- Empirical validation (Brain MRI)
    - Please see the paper for CIFAR-10 results
- Extension to domain adaptation

* For practical guidelines on estimating bounds, please check the paper

Traditional generalization bounds fail to capture the real-synthetic trade-off

- Standard ERM with mixed real + synthetic samples:

$$f_N = \arg\min_f \sum_{n=1}^{N} (y_n - f(x_n))^2 + \sum_{m=1}^{M} (\tilde{y}_m - f(\tilde{x}_m))^2$$

- Problem: classical bounds are too loose. If we apply uniform coverage:

$$\text{Risk} \lesssim \widehat{\mathscr{R}} + \underbrace{\frac{c}{\sqrt{N(1+\lambda)}}}_{\text{Variance}} + \frac{\lambda}{1+\lambda} \underbrace{\text{IPM}(p, p')}_{\text{Bias}}$$

- It suggests a threshold behavior:
  Either use no synthetic data or use unlimited synthetic data

Kernel regression reveals that synthetic data acts as regularization

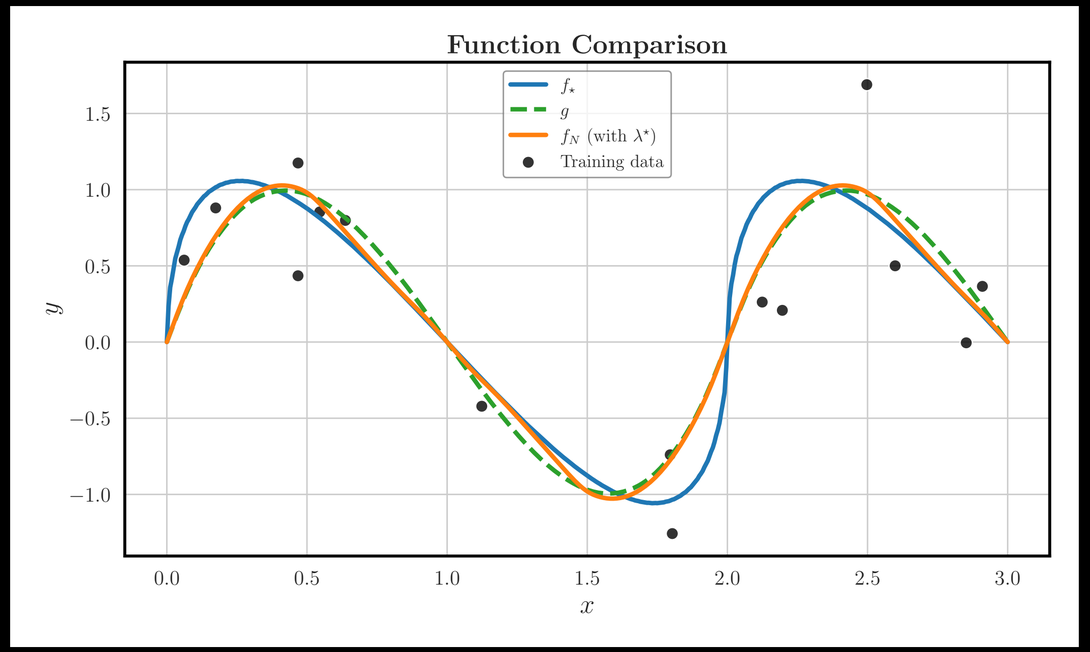
- Interpretation: instead of fitting noisy synthetic samples, regularize toward the synthetic generator g
- This corresponds to having infinite synthetic samples (the population limit)

$$\frac{1-\tilde{\lambda}}{N}\sum(y_n - f(x_n))^2 + \tilde{\lambda}\|f-g\|^2$$

True function: $f_\star$

Synthetic generator: g

Learned function: f_N



An optimal synthetic-to-real ratio exists

And depends on distributional distance

Discrepancy between true function and synthetic generator

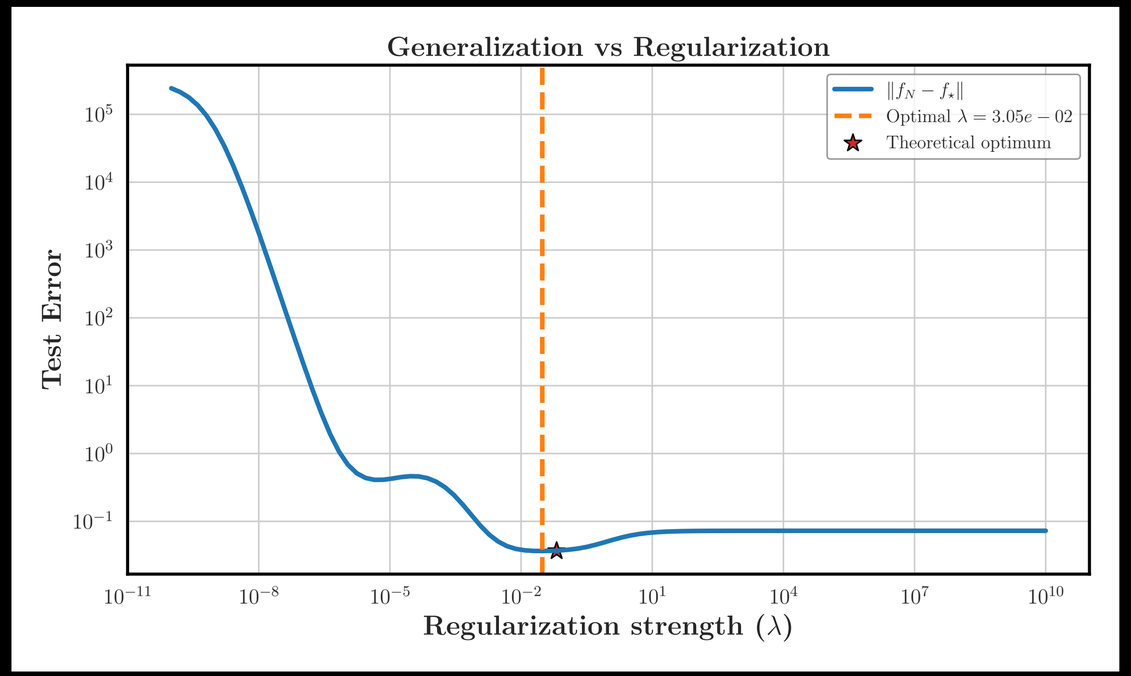
- Theorem (informal):

Test error $$\mathscr{R}_N(\lambda; g) = O\left(\frac{\mathscr{D}(f_\star, g) + \sigma^2}{N\lambda^2} + \lambda^{2-1/(4r)}\mathscr{D}(f_\star, g)\right)$$

- Two competing terms:
  variance decreases with λ
  vs.
  bias increases with λ

- Optimal ratio:

$$\lambda^\star \propto \left(\frac{\mathscr{D}(f_\star, g) + \sigma^2}{N\mathscr{D}(f_\star, g)}\right)^{4r/(16r+1)}$$



The stability framework extends this to general learning algorithms

- Mixed loss:

$$\mathscr{R}_\lambda(h, S) = (1-\lambda)\hat{L}_S(h) + \lambda \mathbb{E}_{x' \sim p'}\left[\ell(h, x')\right]$$

- Tool: Algorithmic stability (uniform stability)
- Key idea: If replacing **one training sample** changes the output **little**, the algorithm generalizes well
- Assumptions: Lipschitz hypothesis class, strongly convex + smooth loss (covers MSE, cross-entropy)
- Wasserstein distance $W_2(p, p')$ replaces $\mathscr{D}(f_\star, g)$

The generalization bound shows a U-shaped dependence on the mixing ratio

- Theorem (Mixed-data Generalization Bound):

$$\mathbb{E}[r(h_S)] \lesssim r^\star + \lambda\xi W_2(p, p') + (1-\lambda) L^{\frac{2d^\star}{d^\star+2}} \Delta^{\frac{2}{d^\star+2}}$$

Irreducible Error

Bias from distributional mismatch

Stability and variance

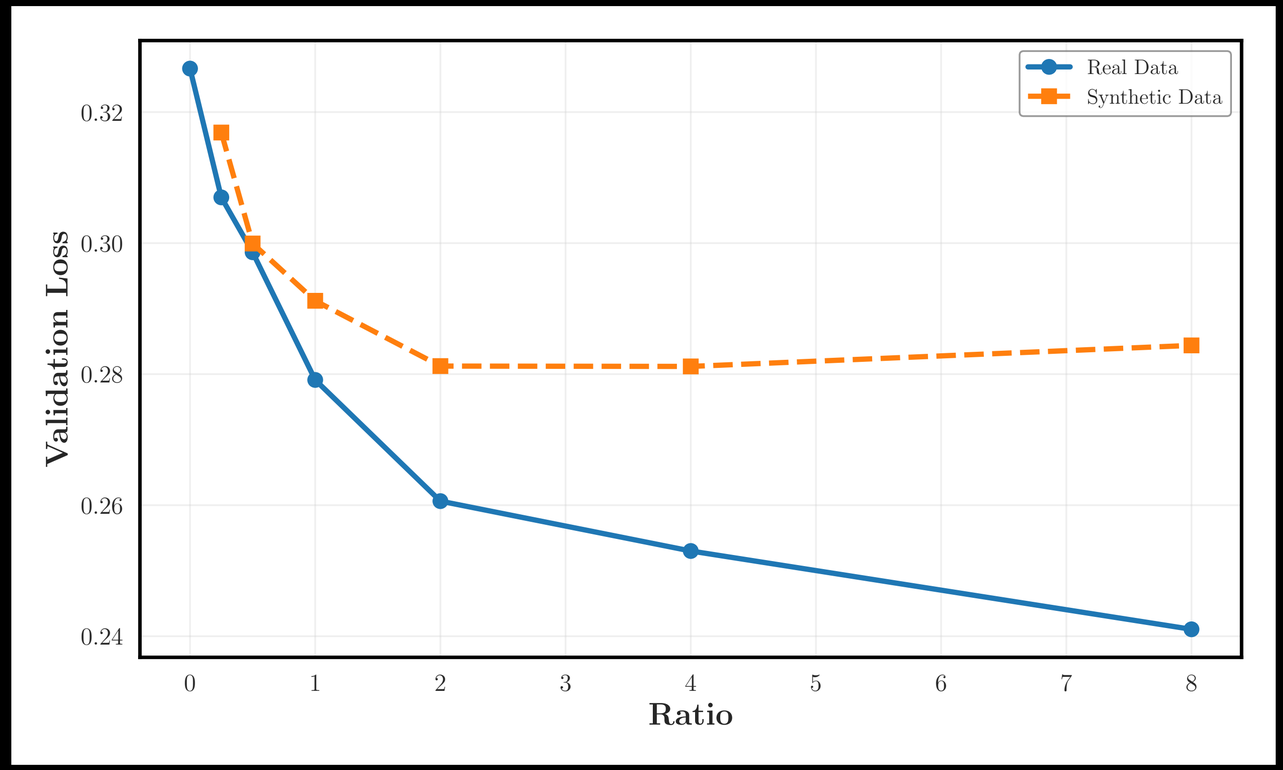
- As $\lambda \uparrow$: bias grows, stability improves
- As $\lambda \downarrow$: stability worsens, bias shrinks
- Special case: $W_2 = 0 \Rightarrow \lambda^\star = 1$ (use as much synthetic data as possible)

Brain MRI experiments confirm the predicted U-shaped behavior

- Dataset: NO.MS — 200K+ MRI scans, MS lesion segmentation
- Setup: 100 real scans + varying amounts of synthetic (ratio 0.25 to 8)
- Synthetic data from conditional diffusion model

Blue (more real data): monotonically improves

Orange (more synthetic data): U-shaped curve

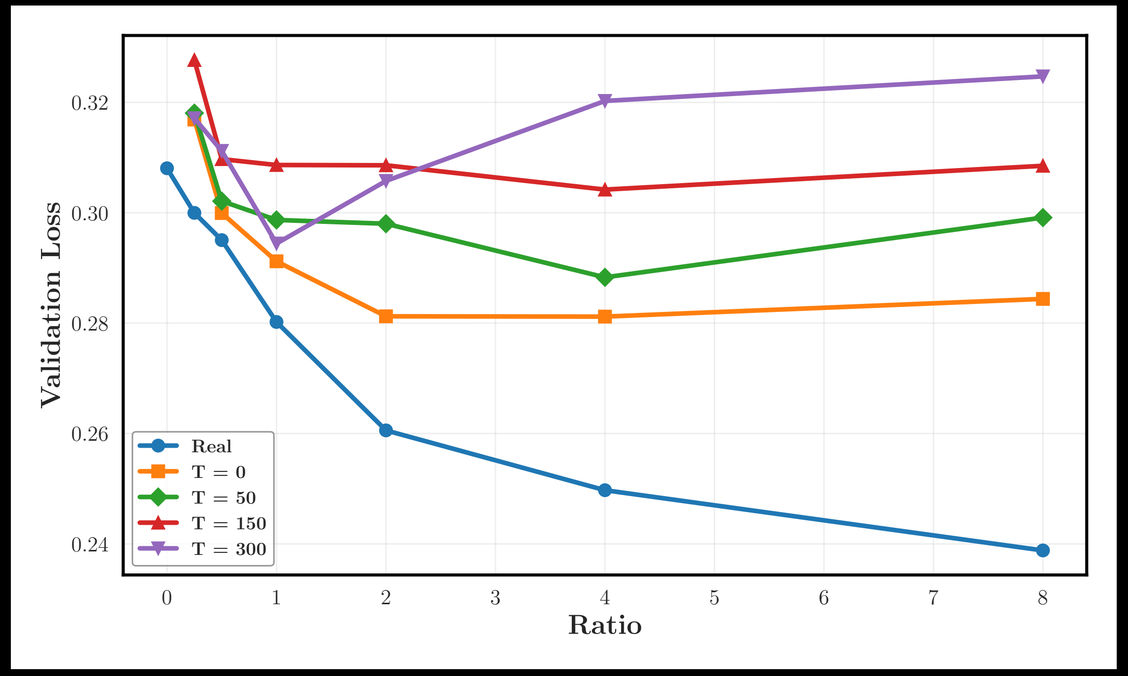


Greater distributional distance shifts and sharpens the optimal ratio

- Varying diffusion time $T \in \{0,50,150,300\}$ as **proxy** for distributional distance
- Higher T = noisier samples = larger distributional distance
- All curves U-shaped, but sharper and shifted left for larger T

Takeaway:

Worse generator quality → need less synthetic data → narrower sweet spot

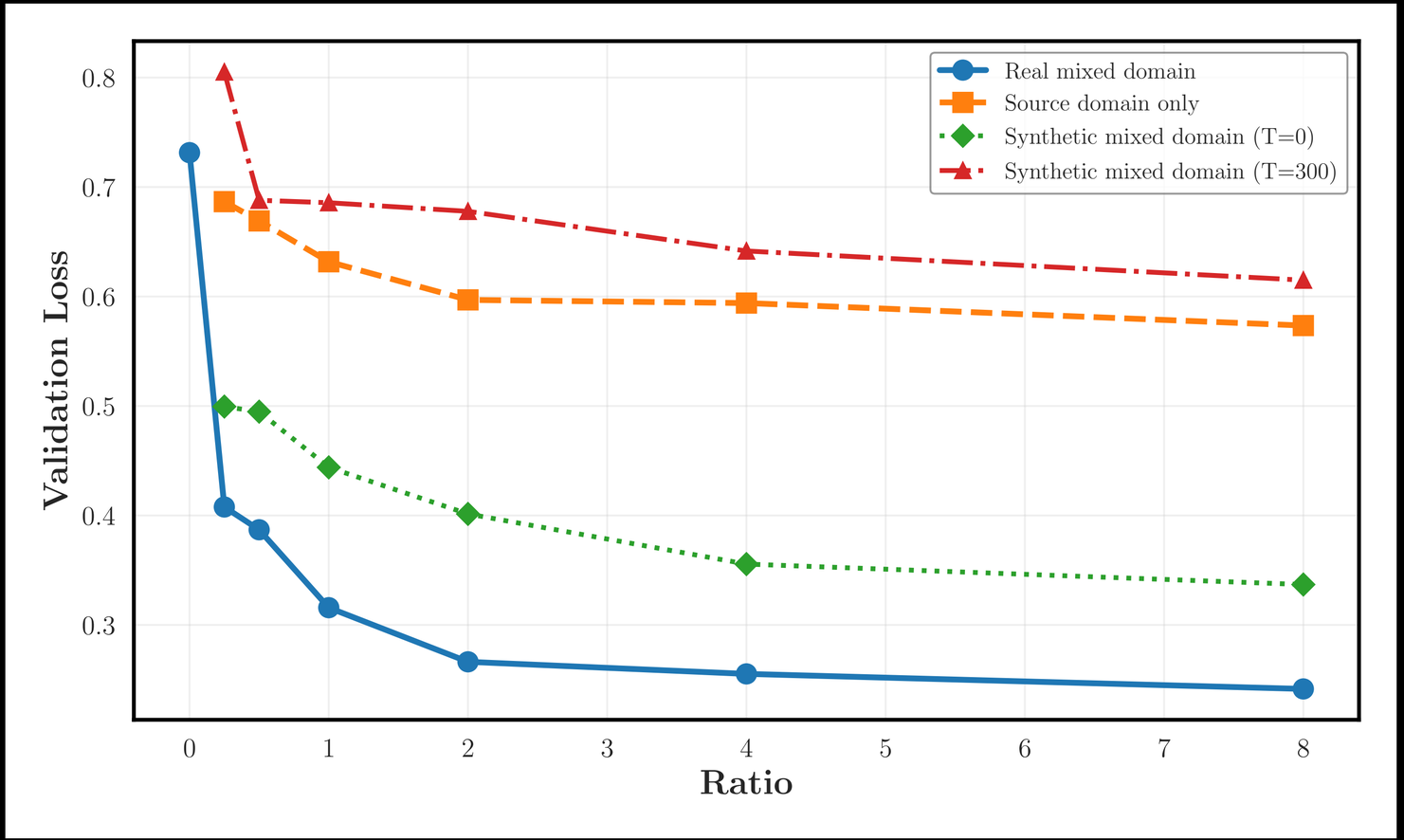


Synthetic data from a target domain can mitigate domain shift

- Setting: Source domain (MIAC) real data + Synthetic data generated on target domain (NeuroRx)
- Extended theorem adds domain shift term:

$$\lambda W_2(p^\star, p') + (1-\lambda) W_2(p^\star, p)$$

- The amount of synthetic data depends on the ratio of distributional distances



Summary: Synthetic data helps, but only with the right balance

- Recap of the story:
  - Synthetic data acts as regularization, not just extra samples
  - An optimal ratio exists, depending on distributional distance, noise, and sample size
  - U-shaped test error confirmed theoretically (kernel + stability) and empirically
  - Extends to domain adaptation: synthetic target data can bridge domain shift

- Open direction: PAC-Bayes bounds treating synthetic data as a prior

Thank You!

If you are interested, please come and see our poster #185