Computational Budget Should Be Considered in Data Selection

Weilin Wan[1], Weizhong Zhang[2], Cheng Jin[1]

1 College of Computer Science and Artificial Intelligence, Fudan University; 2 School of Data Science, Fudan University

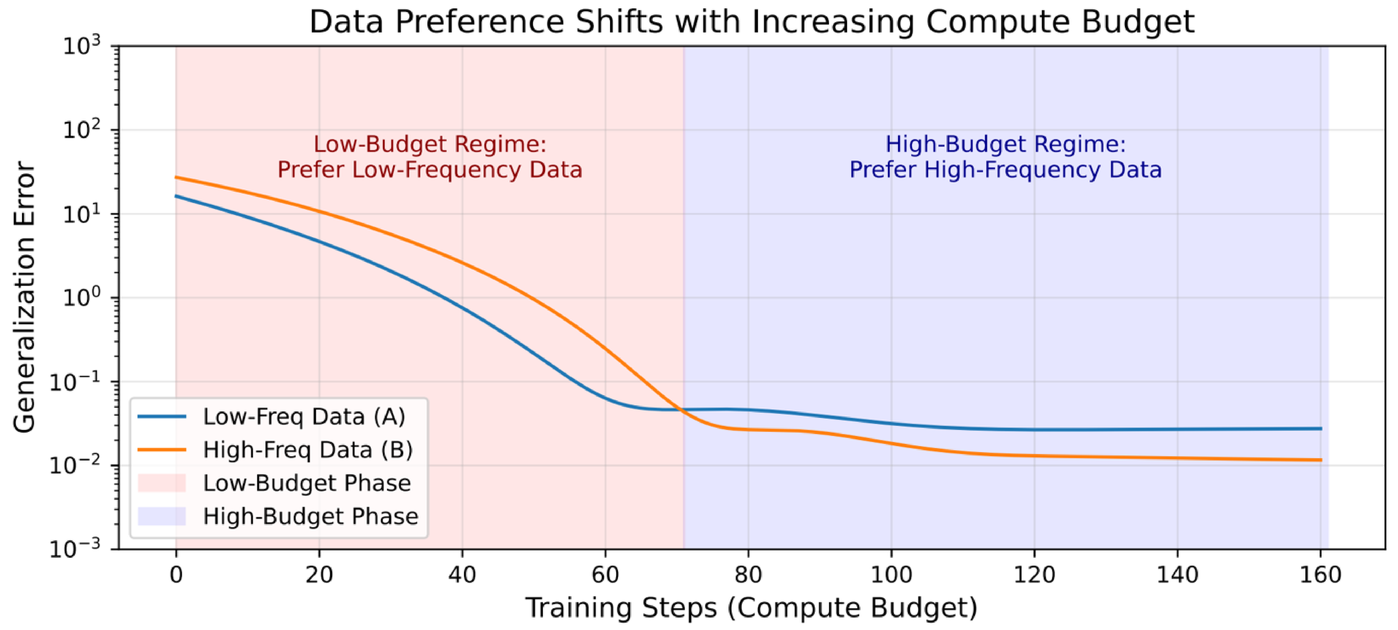
Budgets Shape Data

- We reveal that **compute budget drives data choice**.
- Prior selectors ignore budget, so no method dominates across settings; even random can win.
- We elevate budget to a **first-order design variable**, aligning *quantity*, *quality*, and *distribution* with available FLOPs.

Spectral Bias Example

Our synthetic study reveals how compute budget shapes data preference:

- Low budget favors **low-frequency** data.
- High budget favors **high-frequency** richness.
- This aligns with spectral bias, demanding **budget-aware selection** that adapts as C grows, enabling strategic compute usage.



Bilevel Formulation

We formalize compute-constrained data selection as a bilevel optimization problem:

- Upper Level:

$$\min_{\boldsymbol{m}} \mathcal{L}_{\mathrm{val}}(\boldsymbol{\theta}_C(\boldsymbol{m})) \quad \text{s.t. } \boldsymbol{\theta}_C(\boldsymbol{m}) = \mathrm{Train}(\boldsymbol{m}, C)$$

  Select subset m to minimize validation loss $\mathcal{L}_{val}$.
- Lower Level:

$$\boldsymbol{\theta}_C(\boldsymbol{m}) = \mathrm{Train}(\boldsymbol{m}, C)$$

  Train model parameters $\theta_C(m)$ on subset m with compute budget C.
- Key Difference: Unlike classical bilevel problems solved to convergence, we explicitly constrain training to C steps.

Challenges in Bilevel Optimization

Solving the bilevel problem presents significant challenges:

- Non-Convergence of Inner Problem: Due to the budget constraint, the inner-level training is *not guaranteed to converge*. This invalidates the use of implicit gradient methods.
- High Cost of Policy Gradients: Infeasible implicit gradients necessitate policy gradients, but their reliance on iterative inner-problem training incurs a *prohibitively high computational cost*.

Cracking Bilevel Barriers

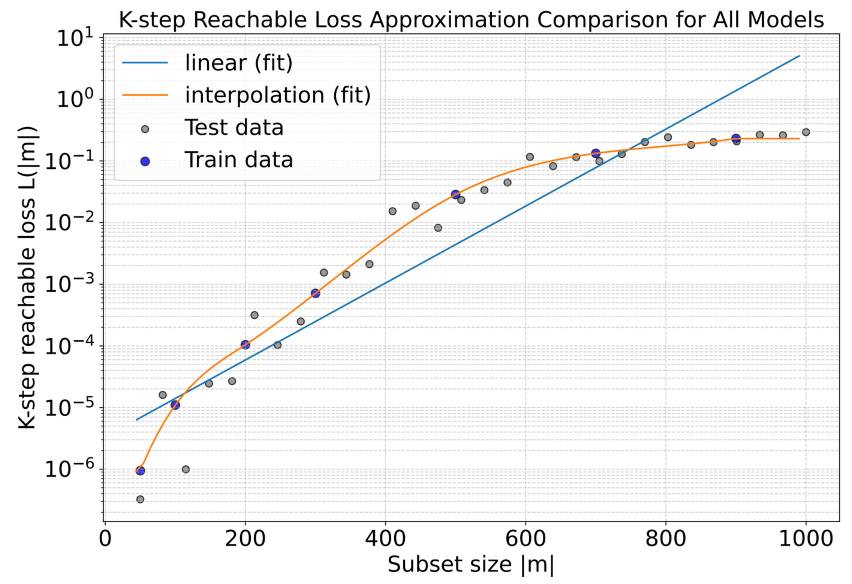
We reformulate it as a single-level problem using a penalty-based relaxation:

$$\min_{\theta,s} \mathcal{L}^{\alpha}_{\text{penalty}}(\theta, s) \triangleq \mathbb{E}_{p(m|s)}\Big[\mathcal{L}_{\mathrm{val}}(\theta) + \alpha\big(\mathcal{L}_{\mathrm{trn}}(\theta, m) - \mathcal{L}_{\mathrm{trn}}(u, m)\big)^2\Big]$$

where $u = \mathrm{Train}(\theta_0, m; K)$ and $K = \frac{C}{|m|}$.

- Challenge with $\mathcal{L}_{\mathrm{trn}}(u, m)$: Obtaining $\mathcal{L}_{\mathrm{trn}}(u, m)$ in the penalty term remains challenging. It is computationally infeasible to retrain u from scratch for every update of m.
- Loss Predictor Approximation: To address this, we approximate $\mathcal{L}_{\mathrm{trn}}(u, m)$ with a loss predictor $l(|m|)$, which estimates the training loss based on the subset size. This yields our final objective:

$$\mathcal{L}^{\alpha}_{\text{CADS}}(\theta, s) = \mathbb{E}_{p(m|s)}\Big[\mathcal{L}_{\mathrm{val}}(\theta) + \alpha\big(\mathcal{L}_{\mathrm{trn}}(\theta, m) - l(|m|)\big)^2\Big]$$



Sampling Points (K)	Mean Square Error
4	0.057668
5	0.016326
6	0.018594
7	0.020031
8	0.017905

MSE of the loss estimator with respect to the number of sampling points (K).

Two Granularities, One Goal

- **CADS-E**: example-level Bernoulli selection for fine-grained control.
- **CADS-S**: source-level truncated-Gaussian ratios with annealed variance, scaling to heterogeneous corpora.

Both variants keep **budget constraints** central to *what* and *how much* we train.

Significant Improvements

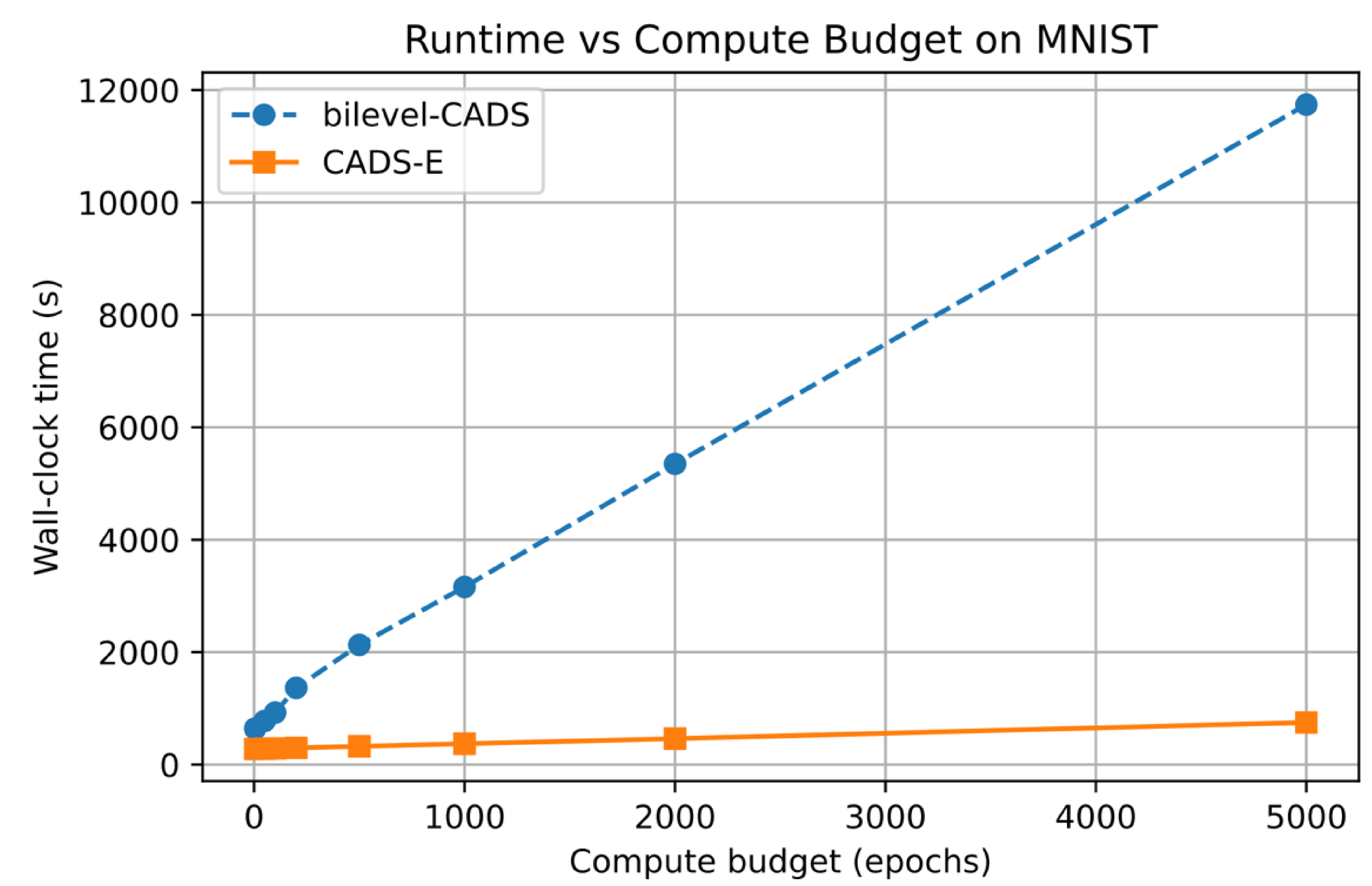
CADS achieves significant performance improvements across vision and language tasks:

- **Significant Accuracy Gains:** Achieves accuracy gains up to 14.42%.
- **Substantial Speedups Achieved:** Delivers substantial speedups of 3–20×.
- **Performance Scales with Budget:** Performance gains **increase with budget**, validating the compute-coupled design.
- **Effective Across Diverse Tasks:** Demonstrated effectiveness on MNIST, CIFAR-10, DomainNet, and instruction tuning.

Efficient Computation

CADS achieves efficient computation through the following techniques:

- **Hessian-Free Operation:** Avoids the use of Hessians.
- **Low Overhead:** Total overhead (5/A + 2 γ) C; with A=5, cost ≈ 2γ **C**.



Limitations and Future Directions

This work highlights several limitations and potential avenues for future research:

- **Algorithm Complexity:** The algorithm's computational complexity remains relatively high, necessitating further optimization.
- **Loss Predictor Dependence:** Performance relies on the accuracy of the loss predictor, which is sensitive to data distribution characteristics.
- **Limited Scope of Budget Consideration:** The current study primarily focuses on compute budget. Future work should explore the impact of parameter size, model architecture, and other resource constraints.