Toward Fast and Provable Data Selection under Low Intrinsic Dimension

Yijun Dong, Per-Gunnar Martinsson, Qi Lei, Hoang Phan, Xiang Pan, Chao Chen, Katherine Pearce

Abstract

As the data volume and model size explode with the unprecedented successes of modern machine learning algorithms, high dimensionality is turning to the major computational bottleneck that impedes the development and democratization of large models. Since redundancies in high dimensions are ubiquitous in most real-world learning problems, the notion of *low intrinsic dimension* is introduced to characterize the minimal size of any low-dimensional manifolds that can encapsulate the essential information in the learning problem. Leveraging such low intrinsic dimensions is crucial for designing fast and sample-efficient learning algorithms for large-scale problems.

Fine-tuning that adapts powerful pre-trained models to specific downstream tasks is arguably one of the most common examples of efficient learning through low intrinsic dimensions. Intuitively, with the general knowledge encoded in the pre-trained model with high-dimensional parameters, fine-tuning within a low-dimensional parameter subspace is usually sufficient for adapting the model to new tasks. Leveraging such low intrinsic dimensions allows learning with much fewer samples (than the high parameter dimension, i.e., in the overparametrized setting) and computational resources.

In practice, natural data generally come with heterogeneous qualities and considerable redundancies, which brings about a critical question:

How to select the most informative data for sample-efficient learning under low intrinsic dimension?

Answers to this question are highly objective-dependent. This talk aims to provide an overview of some recent progress in two common objectives for data selection:

(i) row (or column) subset selection for low-rank interpolative decomposition, and

(ii) data selection for statistical learning models in kernel regime (e.g., fine-tuning).

By diving into a few randomized algorithms for interpolative (or CUR) decompositions and data selection based on random pivoting and sketching, we will unveil the power of randomization in fast and robust data selection, from both the empirical and theoretical perspectives.

Data Selection for Low-rank Interpolative Decompositions

The interpolative decomposition (ID) aims to construct a low-rank approximation formed by a basis consisting of row (or column) skeletons in the original matrix and a corresponding interpolation matrix. We explore fast and accurate ID algorithms from five essential perspectives for empirical performance:

(i) *skeleton complexity* that measures the minimum possible ID rank for a given low-rank approximation error,

(ii) *asymptotic complexity* in floating point operations (FLOPs),

(iii) *parallelizability* of the computational bottleneck, i.e., whether the steps with dominant cost can be cast into matrix-matrix, instead of matrix-vector, multiplications,

(iv) *error-revealing property* that enables automatic rank detection for given error tolerances with-

out prior knowledge of target ranks,

(v) *ID-revealing property* that ensures efficient construction of the optimal interpolation matrix after selecting the skeletons.

While many algorithms have been developed to optimize parts of the aforementioned perspectives, practical ID algorithms proficient in all perspectives remain absent. To fill in the gap, we introduce *robust blockwise random pivoting (RBRP)* that is parallelizable, error-revealing, and exactly ID-revealing, with comparable skeleton and asymptotic complexities to the best existing ID algorithms in practice. Through extensive numerical experiments on various synthetic and natural datasets, we demonstrate the appealing empirical performance of RBRP from the five perspectives above, as well as its robustness to adversarial inputs.

In a nutshell, random pivoting for interpolative decomposition involves adaptively sampling rows (or columns) according to their squared ℓ_2-norm and updating the data matrix by projecting the remaining rows (or columns) onto the orthogonal complement of the current basis. Such an adaptive sampling scheme ensures that the selected rows (or columns) are informative and diverse, leading to a small skeleton complexity for given low-rank approximation errors. However, the sequential nature of random pivoting compromises its parallelizability and empirical efficiency. Alternatively, the sequential random pivoting can be naïvely extended to a faster blockwise version that samples a block of $b > 1$ points according to the current squared ℓ_2-norm in each step and updates the data matrix blockwisely. However, such plain blockwise random pivoting tends to suffer from unnecessarily large skeleton complexity under adversarial inputs due to the lack of local adaptiveness within each block. As a remedy, RBRP leverages *robust blockwise filtering*—applying CPQR to every small sampled block locally and discarding the potentially redundant points through a truncation on the relative residual of the CPQR. By choosing a reasonable block size, such robust blockwise filtering effectively resolves the inefficiency in skeleton complexity encountered by the plain blockwise random pivoting, with negligible additional cost.

Data Selection for Statistical Learning Models in Kernel Regime

Fine-tuning can be viewed as learning with a good pre-trained initialization that lies in some neighborhood of an optimal solution, whose dynamics fall into the kernel regime. Therefore, fine-tuning a regression task (under Tikhonov regularization with a suitable hyperparameter) can be well approximated by

(i) a linear regression problem in the low-dimensional (overdetermined) setting, or

(ii) a ridge regression problem in the high-dimensional (overparametrized) setting[1].

For overdetermined linear regression in low dimension, data selection falls in the classical frames of coreset selection for linear regression and optimal experimental design where the generalization gap can be reduced by selecting data that minimize the associated variance. However, for overparametrized problems, variance minimization alone is insufficient to characterize the generalization. In particular, when the parameter dimension r is higher than the coreset size n, the selected data necessarily miss a parameter subspace of dimension at least $r - n$, leading to errors in addition to variance.

Nevertheless, the prevailing empirical and theoretical evidence on the ubiquitous intrinsic low-dimensional structures in high-dimensional problems motivates a natural question:

[1]We refer to "low-dimension" as the setting where the number of parameters r is smaller than the selected downstream sample size n, while "high-dimension" refers to the opposite, $r > n$.

Can the low intrinsic dimension be leveraged in data selection for high-dimensional fine-tuning?

We provide a positive answer to this question through a variance-bias tradeoff perspective. Intuitively, we consider a low-dimensional subspace $\mathcal{S}$ in the fine-tuning parameter space where the model learns the necessary knowledge for the downstream task. The generalization gap can be controlled by simultaneously reducing *the bias (redundant information) by "exploring" the parameter space to find a suitable* $\mathcal{S}$ and *the variance by "exploiting" the useful knowledge in* $\mathcal{S}$.

Given the high-dimensional nature of the parameter space, a direct search for such suitable subspace $\mathcal{S}$ is computationally infeasible in general. This leads to a follow-up question:

How to explore the intrinsic low-dimensional structure efficiently for data selection?

We propose *Sketchy Moment Matching (SkMM)*, a two-stage solution for this question:

(i) **Gradient sketching for bias reduction**: First, we construct a low-dimensional parameter subspace $\mathcal{S}$ by sketching the model gradients. Sketching is a well-established dimensionality reduction tool known for affordable and accurate low-rank approximations. In deep learning, sketching recently extends its empirical applications to scalable estimations of influence functions for data selection. We make a first step toward the theoretical guarantee of gradient sketching for data selection: *gradient sketching efficiently finds a low-dimensional subspace* $\mathcal{S}$ *with small bias such that selecting* n *samples by reducing variance over* $\mathcal{S}$ *is sufficient to preserve the fast-rate generalization* $O(\dim(\mathcal{S})/n)$*, linear in the low intrinsic dimension* $\dim(\mathcal{S})$ *while independent of the high parameter dimension* r.

(ii) **Moment matching for variance reduction**: Second, we select data that reduce variance in the low-dimensional subspace $\mathcal{S}$ via moment matching. The variance of data selection is characterized by matching between the sketched gradient moments of the original and selected datasets, $\widetilde{\mathbf{\Sigma}}, \widetilde{\mathbf{\Sigma}}_S$, respectively. The objective $\mathrm{tr}(\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}_S^{\dagger})$ takes the form of V-optimality in optimal experimental design, whose exact minimization is computationally intractable. Existing polynomial-time heuristics for V-optimality are generally based on the continuous relaxation of the V-optimality objective followed by a fast rounding process. However, solving such a continuous relaxation can be challenging in practice, as it involves inverting a potentially ill-conditioned matrix $\widetilde{\mathbf{\Sigma}}_S$ in each iteration. Under a common heuristic assumption that $\widetilde{\mathbf{\Sigma}}, \widetilde{\mathbf{\Sigma}}_S$ commute, we introduce a continuous relaxation with a quadratic objective and linear constraints that is numerically stable (free of inversions) and can be efficiently optimized via projected gradient descent.

With synthetic mixtures of Gaussian data, we first demonstrate how SkMM balances variance and bias in data selection for overparametrized ridge regression and leads to sample-efficient learning. Then, with extensive experiments on fine-tuning CLIP or ImageNet pre-trained vision models for both regression and classification tasks, we show the appealing sample and computational efficiency of SkMM, along with its surprising robustness to data heterogeneity.

References

- Robust Blockwise Random Pivoting: Fast and Accurate Adaptive Interpolative Decomposition. Yijun Dong, Chao Chen, Per-Gunnar Martinsson, Katherine Pearce. *arXiv:* 2309.16002, 2023.
- Sketchy Moment Matching: Toward Fast and Provable Data Selection for Finetuning. Yijun Dong*, Hoang Phan*, Xiang Pan*, Qi Lei. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. (to appear)