An Analytical Investigation of Data Sampling Strategies in Stochastic Gradient Descent

Aksel Aghajanyan
Aqwel AI
Armenia
January 2026

1. Introduction

The Role of Data Ordering in Stochastic Optimization

Stochastic Gradient Descent (SGD) is a foundational optimization algorithm in modern machine learning, enabling efficient training of models on large-scale datasets. Despite its algorithmic simplicity, the performance of SGD is highly sensitive to the strategy used to sample training data. This paper presents a formal analytical investigation of two dominant sampling strategies—**uniform sampling with replacement** and **random reshuffling without replacement**—and frames their comparison as a fundamental trade-off between randomness and structure in stochastic optimization.

Rather than treating sampling strategies as a binary choice, this work positions them along a spectrum ranging from pure stochasticity to increasingly structured data orderings. Uniform sampling represents maximal randomness and satisfies classical assumptions of unbiasedness and independence that underpin standard convergence analyses. Random reshuffling introduces structured randomness, violating these assumptions while empirically accelerating convergence. Beyond these two paradigms, even stronger forms of structure arise through deterministic orderings and, in adversarial settings, maliciously constructed data sequences.

The central thesis of this paper is that while random reshuffling typically yields superior convergence behavior, the theoretical basis for this advantage reveals a nuanced interaction between stochasticity, structure, and optimization dynamics. By dissecting the mathematical properties of each sampling strategy, we show why random reshuffling often outperforms uniform sampling despite introducing bias and dependence into gradient estimates. At the same time, we demonstrate that structured randomness is not

universally optimal: carefully chosen deterministic sequences may outperform reshuffling, while adversarial orderings can exploit the same structural sensitivities to disrupt training.

The remainder of this paper is organized as follows. Section 2 reviews the theoretical foundations of SGD and empirical risk minimization. Section 3 formally defines uniform sampling and random reshuffling and highlights their analytical differences. Section 4 presents the theoretical justification for the superior convergence properties of random reshuffling under constant and diminishing step-size regimes. Section 5 broadens the discussion to include deterministic orderings, formal counterexamples, and data ordering attacks. Section 6 examines practical system-level considerations in large-scale machine learning. Section 7 concludes with a synthesis of theoretical, practical, and security-related insights.

---

2. Theoretical Foundations of Stochastic Gradient Descent

A rigorous understanding of data sampling strategies requires first examining the optimization framework within which they operate. SGD is an iterative method designed to solve the empirical risk minimization problem, where the objective is to minimize a loss function aggregated over a finite dataset.

Formally, given a dataset of $n$samples, SGD seeks to minimize

$$L(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta) + R(\theta),$$

where $\ell_i(\theta)$denotes the loss associated with the i-th data point and $R(\theta)$is a regularization term. For large datasets, computing the full gradient of $L(\theta)$at each iteration is computationally prohibitive. SGD addresses this limitation by approximating the gradient using a subset of samples.

Three primary gradient-based optimization methods arise in this context:

- **Full-Batch Gradient Descent**, which computes exact gradients using the entire dataset, yielding stable convergence at high computational cost.
- **Stochastic Gradient Descent**, which uses a single sample per update, achieving low per-iteration cost at the expense of high variance.

- **Mini-Batch Gradient Descent**, which balances computational efficiency and gradient stability by averaging gradients over small batches.

In addition to sampling strategy, the **step-size schedule** critically influences convergence behavior. Constant step-sizes lead to convergence within a neighborhood of the optimum, with the neighborhood size proportional to the step-size. In contrast, diminishing step-size schedules enable convergence to the exact minimizer under suitable conditions.

These foundational elements establish the context in which sampling strategies exert their influence, motivating a detailed comparison of sampling with and without replacement.

---

3. Uniform Sampling and Random Reshuffling

3.1 Uniform Sampling (With Replacement)

Uniform sampling selects each data point independently and uniformly at every iteration. A key theoretical advantage of this strategy is that the resulting stochastic gradient is an unbiased estimator of the true gradient. This property enables tractable convergence analyses and underlies most classical results in stochastic optimization.

However, uniform sampling allows repeated selection of the same data points within a single epoch while potentially excluding others, leading to inefficiencies in data utilization and increased gradient variance.

3.2 Random Reshuffling (Without Replacement)

Random reshuffling generates a random permutation of the dataset at the beginning of each epoch and processes every data point exactly once. This strategy is widely used in practice but introduces two analytical challenges: gradient estimates become biased within an epoch, and successive samples are no longer independent.

Formally, the conditional expectation of the stochastic gradient at iteration $i$within an epoch differs from the true gradient due to the shrinking pool of remaining samples. This violation of classical assumptions complicates theoretical analysis but, as shown in the next section, does not preclude superior convergence behavior.

---

4. Analytical Justification for Random Reshuffling

4.1 Constant Step-Size Regime

Under constant step-sizes and strong convexity assumptions, SGD converges to a steady-state error neighborhood. Uniform sampling yields an error neighborhood of order $O(\mu)$, whereas random reshuffling achieves a substantially smaller neighborhood of order $O(\mu^2)$. This quadratic improvement explains the enhanced precision observed in practice.

4.2 Diminishing Step-Size Regime

With diminishing step-sizes, uniform sampling converges at rate $O(1/i)$. Random reshuffling accelerates this convergence to $O(1/i^2)$, significantly reducing the number of iterations required to achieve a target accuracy.

---

5. The Broader Landscape of Data Ordering

5.1 Deterministic Orderings

Deterministic incremental gradient methods highlight that randomness is not inherently optimal. Carefully constructed deterministic sequences can outperform random reshuffling by exploiting dataset structure.

5.2 Formal Counterexamples

Recent work has disproven conjectures suggesting the universal superiority of reshuffling. Constructed convex problems demonstrate that optimally tuned uniform sampling can converge faster than reshuffling under fixed step-size constraints.

5.3 Data Ordering Attacks

Data ordering also constitutes a security vulnerability. Adversarial batch reordering attacks can degrade model performance or implant backdoors without modifying data or labels, underscoring the sensitivity of SGD to structured sequencing.

---

6. Practical Considerations in Large-Scale Systems

Full data shuffling incurs significant I/O overhead on modern storage systems. Hierarchical strategies such as block-level shuffling provide a practical compromise, preserving statistical efficiency while improving hardware utilization.

---

7. Conclusion

This paper demonstrates that data ordering is a central design choice in stochastic optimization. Random reshuffling offers clear advantages in many regimes, but its superiority is conditional rather than universal. Deterministic and adversarial structures reveal both opportunities and risks, positioning data ordering as a powerful yet delicate lever in modern machine learning systems. A principled understanding of this trade-off is essential for developing robust, efficient, and secure optimization pipelines.

Author Contributions:

Aksel Aghajanyan conceived the research problem, conducted the theoretical analysis, and authored the manuscript. The work was carried out independently at Aqwel AI, where the author serves as the lead researcher.